# From transposon to transcription factor:

# Genome-wide functional studies of the conserved primate CSB-PGBD3 fusion protein

Lucas Gray

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2012

Reading Committee:

Alan Weiner, Chair

Nancy Maizels

Jay Shendure

Program Authorized to Offer Degree:

Department of Biochemistry

University of Washington

## Abstract

From transposon to transcription factor:

Genome-wide functional studies of the conserved primate CSB-PGBD3 fusion protein

Lucas Gray

Chair of the Supervisory Committee:
Dr. Alan Weiner PhD.
Department of Biochemistry

The CSB-PGBD3 fusion protein arose over 43 million years ago when a 2.5 kb piggyBac 3 (PGBD3) transposon inserted into intron 5 of the Cockayne syndrome Group B (CSB) gene in the common ancestor of all higher primates. As a result, full length CSB is now coexpressed with an abundant CSB-PGBD3 fusion protein by alternative splicing of CSB exons 1-5 to the PGBD3 transposase. An internal deletion of the piggyBac transposase ORF also gave rise to 889 dispersed, 140 bp MER85 elements which were mobilized in trans by PGBD3 transposase. Here, we show that the CSB-PGBD3 fusion protein binds MER85s *in vitro*, and induces a strong interferon-like innate antiviral immune response when expressed in CSB-null UVSS1KO cells. To explore the connection between DNA binding and gene expression changes induced by CSB-PGBD3, we investigated the genome-wide DNA binding profile of the fusion protein. CSB-PGBD3 binds to 363 MER85 elements *in vivo*, but these sites do not correlate with gene expression changes induced by the fusion protein. Instead, CSB-PGBD3 is enriched at AP-1, TEAD1, and CTCF motifs, presumably through protein-protein interactions with the cognate transcription factors; moreover, recruitment of

CSB-PGBD3 to AP-1 and TEAD1 motifs correlates with nearby genes regulated by CSB-PGBD3 expression in UVSS1KO cells and downregulated by CSB rescue of mutant CS1AN cells. We also examined close PGBD3 homologs in galago monkeys, which have no domesticated PGBD3, and the freshwater cnidarian *Hydra magnipapillata* to show that as many as 30 mutations lead to the domestication of PGBD3. We conclude that horizontal transfer of PGBD3 created the CSB-PGBD3 fusion protein, which substantially reshapes the transcriptome in CS patient CS1AN, and that continued expression of the CSB-PGBD3 fusion protein in the absence of functional CSB may affect the clinical presentation of CS patients by directly altering the transcriptional program.

# Table of Contents

## List of Figures

## List of Tables

# Acknowledgments

First and foremost, my thanks to my advisor Alan Weiner, who guided this research, spent countless hours poring over presentations, posters, and prose, and allowed me to pursue new bioinformatic and analytical techniques. I feel I have grown immensely over the course of the last six years as a scientist, author, presenter, and analyst thanks to our frequent discussions.

Thanks also go to my coworkers and coauthors Arnold Bailey and Thomas Pavelitz, without whose technical expertise and training my project would not have been successful. Arnold and Tom made many of the plasmid constructs and cell lines that were essential for the experiments that are described in the following chapters, and taught me many biochemical and cell culture techniques that proved invaluable. In addition, Arnold purified RNA and performed the initial expression array analysis that was the basis for Chapter 2, and Tom performed gel shift assays that are in both Chapter 2 and 3. Kimberly Fong did the QPCR analysis used in Chapter 3.

I also thank my thesis committee, Harmit Malik, Ray Monnat, Jay Shendure, Nancy Maizels, and Brian Kennedy, who have each provided valuable discussion and suggestions that have helped me to pursue my research goals.

Jay Shendure and Charlie Lee in the UW Department of Genome Sciences provided the expertise and equipment that made my high throughput sequencing experiments possible. These experiments were a cornerstone of my thesis work, and I am thankful for their assistance.

This work was funded by the Cell and Molecular Biology Training Grant Program T32 GM007270 (LTG) and NIH award R01 GM41624 (Alan Weiner).

# Dedication

To my parents
Mark and Sheila Gray

# Chapter 1:
# What is PGBD3 doing in our genome?

### Cockayne syndrome: at the intersection of transcription, repair, chromatin remodeling, and ubiquitylation

Genomic DNA is under constant mutational stress. Maintaining the sequence of the genome involves many complex processes specific to different types of damage. To keep genes from being mutated or disabled, transcribed regions are protected by a specialized repair process called transcription-coupled nucleotide excision repair (TC-NER). First, DNA damage is recognized by RNA polymerase II (RNAPII), which stalls when it encounters helix-distorting lesions on the template strand such as pyrimidine dimers caused by UV light. The stalled polymerase is then recognized by the Cockayne syndrome group B (CSB) protein, which recruits NER factors, including TFIIH, several xeroderma pigmentosum (XP) proteins, and the Cullin-DDB1-CSA ubiquitylation complex [1]. Transcription resumes after the lesion is repaired.

Loss of either CSA or CSB function results in the severe autosomal recessive disorder Cockayne syndrome, in which patients exhibit accelerated aging, neurodevelopmental dysfunction, and UV sensitivity [2]. The reason that loss of CSA or CSB function causes the same phenotype has been quite mysterious. CSB is a SWI/SNF2 DNA-dependent ATPase protein that plays a role in elongation by RNAPI [3] and RNAPII [4], mediates the interaction between the transcription factors p300 and p53 [5], remodels chromatin structure in the absence of DNA damage [6], and recruits repair factors to DNA lesions on the template strand where RNAPII has stalled. CSA, on the other hand, is a ubiquitin ligase that is recruited to the site of DNA damage after CSB. In contrast, CSA is able to ubiquitylate CSB at a late stage in the repair process, resulting in degradation of CSB by the proteasome pathway and resumption of RNA synthesis [7]. Thus CSA may be necessary for successful DNA repair, or for disassembly of the TCR-NER complex during repair or afterwards. In fact, when CSB lacks its ubiquitin binding domain, TCR complexes fully assemble, but no repair occurs [8]; also see Preview by Gray, LT and Weiner, AM [9]. Instead, CSB and the entire TC-NER complex appear to be trapped at sites of UV damage, sequestering CSB and preventing it from participating in other functions such as transcription initiation, elongation, and chromatin remodeling.

Thus loss of CSA or CSB function may cause clinically indistinguishable forms of CS because loss of CSB affects TCR, chromatin remodeling, and transcription elongation directly, whereas loss of CSA does so indirectly by sequestering CSB in stalled TCR complexes.

Sequestration of CSB at the site of stalled RNAPII could also explain the puzzling

observation that loss of any of the 7 xeroderma pigmentosum factors (XPA, B, C, D, E, F and G) causes cancer-prone cutaneous UV sensitivity, whereas only rare alleles of XPB, D, and G cause CS. Unlike transcription-coupled nucleotide excision repair (TC-NER or TCR) which repairs only the template strand in transcribed DNA, global genome NER (GG-NER or GGR) repairs both DNA strands in nontranscribed DNA. In contrast to GGR of UV damage which requires all 7 factors (XPA, B, C, D, E, F and G), TC-NER requires only 5 of these factors (XPA, B, D, F and G) because the roles of XPC and XPE in recognizing and partially unwinding DNA damage are performed by RNAPII in TC-NER. Thus CSB and CSA can be thought of as assembly factors for building an NER complex at sites where RNAPII has stalled.

An attractive hypothesis would then be that those rare alleles of XPB, D, and G that cause CS do so by allowing assembly of a stable but nonfunctional TC-NER complex that sequesters CSB, whereas the vast majority of mutations in the 7 XP factors either prevent formation of a stable TC-NER complex, or fail to repair the DNA lesion after release of CSB and thus cannot stably sequester CSB at the site. This hypothesis would further imply that the accumulation of mutations normally repaired by NER plays a secondary role in CS, and that CSB cannot be sequestered by stalled RNAPII without further stabilizing interactions with XP factors. Indeed, Ito et al. (2007) have demonstrated that XPG is required for stabilization of the TFIIH complex containing XPB, XPD, and XPG [10]. Similarly, loss of the UVSSA protein, which stabilizes CSB after high doses of UV damage [11-13], would not cause CS because ongoing synthesis of CSB would compensate for loss of CSB engaged in TC-NER.

**A role for the CSB-PGBD3 fusion protein in Cockayne syndrome?**

While examining expressed sequence tags (ESTs) for human CSB, John Newman of our research group discovered an unannotated, alternatively spliced form of CSB that joined the first five exons of the CSB gene in frame to a transposase ORF within intron 5 of the CSB gene. This transposase is part of piggyBac derived element 3 (PGBD3), a DNA transposon encoding a cut and paste transposase which recognizes the terminal inverted repeats of the element. John discovered several aspects of the CSB-PGBD3 protein that would form the basis of my project. He found that the transposase ORF is flanked immediately upstream by a 3' splice acceptor site in frame with CSB exon 5, and downstream by a functional polyadenylation site. Together, these two RNA processing signals allow the PGBD3 transposase ORF to function as an alternative 3' terminal exon, thus generating a chimeric protein in which the N-terminal CSB domain is fused to the C-terminal PGBD3 transposase. He demonstrated that the fusion protein is expressed in human cells, including those of several Cockayne syndrome patients with CSB mutations downstream of the boundary between exons 5 and 6. He showed that the PGBD3 insertion in CSB had been conserved in humans, chimpanzees, gorillas,

and marmosets going back at least 43 million years (My); he realized that a cryptic promoter in CSB exon 5 allowed independent expression of the PGBD3 transposase; and he recognized that PGBD3 had mobilized hundreds of nonautonomous elements, called MER85s, which consist of the first 100 bp and the last 40 bp of the PGBD3 transposon. And finally, John found these MER85 elements were enriched near genes down-regulated by UV irradiation or expression of CSB in a CS cell line [14].

Although it is possible that the CSB-PGBD3 fusion protein has nothing to do with the causes, clinical heterogeneity, or severity of Cockayne syndrome, it is difficult to believe that a highly conserved fusion protein which bears the acidic N-terminal domain of the CSB protein and is coexpressed with CSB would *not* play some role in Cockayne syndrome. At the time I joined the research group, it was known that most CS patients with molecularly characterized mutations in the CSB gene continued to express the CSB-PGBD3 fusion protein from at least one allele. The only patient with a total knockout of both CSB and CSB-PGBD3 exhibited UV Sensitivity syndrome (UVSS), but no other symptoms of CS [15], suggesting that loss of both CSB and CSB-PGBD3 was less detrimental than expression of CSB-PGBD3 alone. Mouse models of CS were also potentially consistent with a role for CSB-PGBD3 in CS, because mice do not express CSB-PGBD3, and a mouse CSB knockout does not cause a CS phenotype [16]]. Instead, a additional DNA repair gene (XPA [17] or XPC [18]) must be knocked out as well in order to generate a phenotype that approximates the disease as seen in human patients.

Though the stage had been set by John's careful research for further studies of CSB function, I chose to work on the CSB-PGBD3 fusion protein because it was a project full of mysteries. What was a piggyBac transposase, a selfish genetic element, doing in the middle of the CSB gene? And why would this invader not only persist, but be fixed and then conserved in the three anthropoid lineages for 43 million years? And if CSB activity was lost by mutation or sequestration (as described above), could CSB-PGBD3 contribute to Cockayne syndrome by interfering with processes normally performed by CSB?

These questions grew even more tantalizing as I learned that nearly half of our genome is composed of mobile elements. Although RNA retroposons like SINEs and LINEs account for the vast majority of these insertions, DNA transposons like PGBD3 occupy about 3% of the genome — over twice as much as coding exons [19]. However, most insertions decay, either because they are detrimental, or because they are neutral and lost to mutation. This was the fate that befell the other 4 full-length PGBD3 insertions, which are now pseudogenes full of frame shifts and stop codons. Indeed, CSB-PGBD3 could easily have been disabled by a single point mutation in the transposase ORF or

the 3' splice acceptor site. And yet it persists.

**Progress in understanding CSB-PGBD3 function**

Over the course of my doctoral studies, several papers substantially expanded the known roles of domesticated transposons in eukaryotic genomes including: the recognition that many transcription factors were originally transposase proteins [20]; the rewiring of the p53 regulatory network in primates by mobilization of endogenous retroviral elements containing p53 binding sites [21]; the creation of new gene networks in pregnancy by mobilization of MER20 elements containing clusters of transcription factor binding sites [22]; and an essential role for the PGBD3-related piggyMac transposase in the programmed genomic rearrangements of the cilliate Paramecium tetraurelia [23]. Thus, transposons do indeed perform many domestic regulatory functions as McClintock [24] and Britten and Davidson [25] foresaw many years earlier, but the mode and tempo of domestication are not well understood. A detailed study of the CSB-PGBD3 fusion protein struck me as an ideal experimental tool for addressing these questions. The fusion protein was old enough to be clearly conserved from marmoset to human, but young enough that both autonomous and nonautonomous members of the PGBD3 transposon superfamily were still recognizable throughout the genome.

I hypothesized that CSB-PGBD3 might have created a new regulatory network by binding to dispersed MER85 elements, and could contribute to CS by interfering with complexes that normally associate with CSB. What I found, however, was both more complicated and more interesting, as the following chapters attest. To explore my hypotheses, I combined a systems biology approach for analysis of genome-wide expression changes with a biochemical approach for protein-DNA interactions.

In Chapter 2, I present the results of an extensive study of the effects of CSB and CSB-PGBD3 on gene expression in the CSB-null cell line UVSS1KO. This patient-derived cell line expresses neither CSB nor CSB-PGBD3, which allowed us to observe changes in gene expression caused by expression of CSB, CSB-PGBD3, or both proteins in a CSB-null background. Using gene set enrichment analysis (GSEA), I found that CSB-PGBD3 triggers expression of an innate immune response that is very similar to the set of genes upregulated by the unphosphorylated forms of STAT1 and STAT2 during the later stages of the interferon response. Through internal comparisons of our expression array results, I found that CSB-PGBD3 and CSB have both antagonistic and cooperative roles in gene regulation. Many genes activated by CSB-PGBD3 expression, including those of the inflammatory response, are suppressed by coexpression of CSB with CSB-PGBD3, and expression of both proteins activates a set of genes that is distinct from those regulated by either CSB or CSB-PGBD3 alone. Chapter 2 was

originally published in DNA Repair [26].

Though expression analysis provided insight into the biological roles of CSB-PGBD3, it provided little in the way of mechanism because expression arrays cannot distinguish between primary, secondary, and more indirect effects of transcription factors. I therefore sought to determine whether the observed gene expresion changes correlated with the presence of a nearby genomic binding sites for the CSB-PGBD3 fusion protein. I performed chromatin immunoprecipitation of CSB-PGBD3 coupled to massively parallel DNA sequencing (ChIP-seq) to generate a genome-wide binding profile for CSB-PGBD3 in the same UVSS1KO cell line used for our expression array analysis. In Chapter 3, I show that CSB-PGBD3 is strongly enriched at more than 2,000 sites in the human genome, including 367 of the 889 MER85 elements. Surprisingly, I found that CSB-PGBD3 also interacts with at least 3 transcription factors that long predate PGBD3 transposition — c-Jun, TEAD1, and CTCF — and demonstrated that CSB-PGBD3 interacts directly with c-Jun and RNAPII through the N-terminal CSB domain. The binding of CSB-PGBD3 to c-Jun correlates with expression changes induced by CSB-PGBD3, while binding of CSB-PGBD3 to MER85 elements did not. Chapter 3 will be published in PLoS Genetics (LT Gray, KK Fong, T Pavelitz, and AM Weiner (2012) PLoS Genetics 8, in press).

In Chapter 4, I examine the origins of PGBD3 and MER85 in primate genomes. The genome of the bushbaby (*Otolemur garnetii*) has many copies of decayed PGBD3 elements, but no PGBD3 insertion in CSB. Thus, this genome shows us the fate of PGBD3 in a context in which PGBD3 is never domesticated. I also discuss the surprising finding that PGBD3 has a direct homolog in the freshwater Cnidarian *Hydra magnipapillata*, providing strong evidence for the horizontal transfer of PGBD3.

While my work has characterized and correlated the transcriptional response to CSB-PGBD3 expression with the genomic DNA binding profile of the fusion protein in CSB-null UVSS1KO cells, these experiments have raised new questions about the mechanism of CSB-PGBD3 in gene regulation, the role CSB-PGBD3 plays in normal cells and CS patients, the regulation of CSB-PGBD3 by CSB, and the evolutionary origins of the CSB-PGBD3 fusion protein. In Chapter 5, I discuss the major current questions regarding CSB-PGBD3 biology, and suggest further research to address these questions based on my results.

# Chapter 2:
# The conserved Cockayne syndrome B-piggyBac fusion protein (CSB-PGBD3) affects DNA repair and induces both interferon-like and innate antiviral responses in CSB-null cells

## Summary

Cockayne syndrome is a segmental progeria most often caused by mutations in the *CSB* gene encoding a SWI/SNF-like ATPase required for transcription-coupled DNA repair (TCR). Over 43 Mya, before marmosets diverged from humans, a piggyBac3 (PGBD3) transposable element integrated into intron 5 of the CSB gene. As a result, primate *CSB* genes now generate both CSB protein and a conserved CSB-PGBD3 fusion protein in which the first 5 exons of CSB are alternatively spliced to the PGBD3 transposase. Using a host cell reactivation assay, we show that the fusion protein inhibits TCR of oxidative damage but facilitates TCR of UV damage. We also show by microarray analysis that expression of the fusion protein alone in CSB-null UV-sensitive syndrome (UVSS) cells induces an interferon-like response that resembles both the innate antiviral response and the prolonged interferon response normally maintained by unphosphorylated STAT1 (U-STAT1); moreover, as might be expected based on conservation of the fusion protein, this potentially cytotoxic interferon-like response is largely reversed by coexpression of functional CSB protein. Interestingly, expression of CSB and the CSB-PGBD3 fusion protein together, but neither alone, upregulates the insulin growth factor binding protein IGFBP5 and downregulates IGFBP7, suggesting that the fusion protein may also confer a metabolic advantage, perhaps in the presence of DNA damage. Finally, we show that the fusion protein binds in vitro to members of a dispersed family of 900 internally deleted piggyBac elements known as MER85s, providing a potential mechanism by which the fusion protein could exert widespread effects on gene expression. Our data suggest that the CSB-PGBD3 fusion protein is important in both health and disease, and could play a role in Cockayne syndrome.

# Introduction

Cockayne syndrome (CS) is a devastating and ultimately fatal progeroid syndrome affecting hundreds of children and occasional adults worldwide. Although often apparently normal at birth, children affected by this multisystem disorder soon exhibit postnatal growth failure, wasting (cachexia), progressive neurological and retinal degeneration, mental retardation, skeletal abnormalities, gait defects, and sun sensitivity, but never an increase in skin cancer or other tumors.

Most cases of CS are caused by mutations in the Cockayne syndrome Group B gene (*CSB*, also known as *ERCC6*) encoding a SWI/SNF-like DNA-dependent ATPase that can wind DNA [27] and remodel chromatin both in vitro [28] and in vivo [6]. The remaining cases of CS are caused by mutations in the *CSA* gene [29] which is required for ubiquitin-dependent degradation of CSB [7,8,30], or by rare alleles of the xeroderma pigmentosum (XP) genes *XPB*, *XPD*, and *XPG* [2]. These three XP genes are required along with *XPA*, *XPC*, *XPE*, and *XPF* for nucleotide excision repair (NER), and loss of XP gene function results in susceptibility to skin cancer.

Significantly, all 5 genes that can cause CS (*CSA*, *CSB*, *XPB*, *XPD*, and *XPG*) are required for transcription-coupled repair (TCR or TC-NER) where XPB and XPD are subunits of TFIIH and XPG is required to stabilize TFIIH [10]. In TCR, actively transcribing RNA polymerase II (pol II) stalls at DNA damage, triggering assembly of an NER complex that repairs the transcribed strand of the DNA and allows transcription to proceed [31]. TCR is distinct from global genome repair (GGR) which detects and repairs DNA damage on both strands of the DNA independently of transcription throughout the cell cycle. Although the NER complexes formed in TCR and GGR contain the same core factors (XPA, XPB, XPD, XPF and XPG), the GGR complex requires two additional proteins (XPC and XPE) whose functions in recognizing and partially unwinding the DNA damage are performed by pol II in TCR. An emerging view is that CSB serves as an adaptor to assemble a stable NER complex wherever pol II has stalled at DNA damage, and CSA then removes CSB and pol II leaving an NER complex in place [8,9]. Thus *CSA*, *CSB*, *XPB*, *XPD*, and *XPG* mutations that cause CS may do so not just *directly* by failing to carry out TCR, but also *indirectly* by trapping scarce CSB in stable nonfunctional TCR complexes; the resulting depletion of free CSB could then affect many genes whose transcription or chromatin structure is dependent on CSB in normal growth [6] and in hypoxia [5].

Although CS is usually recessive, complete loss of CSB function does not invariably cause CS. A 33 year old male, UVSS1KO, who expressed no CSB-related proteins as a

result of a nonsense mutation at *CSB* codon 77, exhibited UV sensitive syndrome (UVSS) but no other CS symptoms [15]. A 47 year old woman, KPSX6, with a frameshift mutation at the same codon, was initially diagnosed with UVSS and did not exhibit late-onset progeria until age 45 [32]. Thus the complete absence of CSB protein can in fact be less harmful than expression of larger CSB nonsense fragments or full length missense mutants. Most recently, Laugel et al. [33] described two CS patients, CS539VI and CS548VI, in which identical homozygous mutations spanning the 5' UTR eliminate all CSB transcription, yet cause classical early-onset CS. Interestingly, all four of these unusual UVSS or CS individuals with complete loss of CSB expression appear to be consanguineous: The parents of UVSS1KO are first cousins; the parents of KPSX6 are said to be consanguineous; and patients CS539VI and CS548VI, although apparently unrelated, are both from the highly inbred population of Reunion Island consistent with a founder effect. Consanguinity in all four of these cases may not be coincidental, and suggests that genetic background might delay or accelerate the appearance of CS symptoms. Indeed, background effects could explain the unusually heterogeneous onset, severity, and multiplicity of CS symptoms [34] as well as the telling observation that the same *CSB* R735opal mutation can cause either CS or a form of XP known as DeSanctis-Cacchione syndrome [35].

Several years ago, we found that the piggyBac transposable element PGBD3 had integrated into intron 5 of the primate CSB gene before marmosets diverged from humans >43 Mya [14]. As a result, the primate *CSB* gene now generates three proteins: intact CSB, a more abundant CSB-PGBD3 fusion protein in which the first 5 of the 22 CSB exons are alternatively spliced to the PGBD3 transposase, and most abundant of all, solitary PGBD3 transposase transcribed from an internal promoter in CSB exon 5 (Figure 2-1). Conservation of the CSB-PGBD3 fusion protein for >43 My strongly suggests that the fusion protein is advantageous in the presence of functional CSB; and the shared N-terminal CSB domain suggests that CSB and the CSB-PGBD3 fusion protein may be functionally related.

*CSB* mutations that cause CS are uniformly distributed over the entire *CSB* coding region [2] but only those nonsense and frameshift mutants located downstream of intron 5 continue to make the CSB-PGBD3 fusion protein ([14] and Figure 2-1). The implication is that CS usually reflects loss of functional CSB but does not require, and may even be unaffected by, continued expression of the CSB-PGBD3 fusion protein. A priori, however, the CSB-PGBD3 fusion protein could be advantageous, neutral, or disadvantageous in the *absence* of functional CSB.

In order to understand the roles of the CSB-PGBD3 fusion protein in health and disease, we set out to investigate the functions of the protein experimentally. We show here that (1) the fusion protein inhibits TCR of oxidative damage but facilitates TCR of UV damage — demonstrating that it can modulate DNA damage responses; (2) expression of the fusion protein in CSB-null UV-sensitive syndrome (UVSS) cells induces an interferon-like response resembling both the innate antiviral response as well as the prolonged interferon response normally maintained by unphosphorylated STAT1 (U-STAT1) — implying that the fusion protein may elevate basal levels of antiviral and antipathogen defenses; (3) coexpression of the fusion protein with CSB upregulates the insulin growth factor binding protein IGFBP5 and downregulates IGFBP7 — suggesting that the fusion protein may also confer a metabolic advantage; and finally (4) the fusion protein binds in vitro to a dispersed family of 900 internally deleted piggyBac elements known as MER85s — suggesting that the CSB-PGD3 fusion protein, when bound to MER85 or related elements, may regulate expression of nearby genes. Taken together, our data support the hypothesis that the CSB-PGBD3 fusion protein is important in health, and may also play a role in CS disease.

# Results

### DNA repair assays

We assayed the effects of the CSB-PGBD3 fusion protein on DNA repair in CSB-null UVSS1KO cells using both a recovery of RNA synthesis (RRS) assay after UV irradiation of whole cells (Figure 2-2A), and host cell reactivation (HCR) assays after UV irradiation (Figure 2-2B) or osmium tetroxide oxidation (Figure 2-2C) of an EGFP reporter plasmid prior to transfection. Importantly, osmium tetroxide ($OsO_4$) oxidation generates thymine glycol damage, which is known to require CSB for efficient repair [36].

In the UVSS1KO-derived cells used for the RRS and HCR assays (Table 2-1), CSB levels were 2- to 4-fold higher and CSB-PGBD3 fusion protein levels 4- to 8-fold higher than in human euploid HT1080 fibrosarcoma cells as determined by Western blots of a dilution series probed for the FLAG-HA tag (Figure 2-3). We have consistently found that the natural ratio of CSB to CSB-PGBD3 fusion protein is about 1 to 4 in many human cell lines, whether the line expresses high levels of CSB (HT1080) or 8- to 10-fold lower levels (WI38). Thus the ratio of CSB to CSB-PGBD3 fusion protein is within the normal range for the cells used in the RRS and HCR assays. We deliberately did not resort to differential siRNA knockdown of CSB and CSB-PGBD3 in normal cells because knockdown is never complete [37], very low levels of CSB expression do not

impair normal RRS in WI38 cells ([6]; J. C. Newman and A. D. Bailey, unpublished observations), and the CSB-PGBD3 fusion protein cannot be efficiently knocked down by siRNA without simultaneously knocking down CSB and/or solitary PGBD3 transposase (Figure 2-1B).

Surprisingly, expression of the CSB-PGBD3 fusion protein alone was almost 40% as effective as expression of intact CSB protein in restoring RRS (Figure 2-2A) although the fusion protein lacks all of CSB's conserved ATPase motifs (Figure 2-1B and [14]). Although it is difficult to imagine a direct role for the CSB-PGBD3 fusion protein in repair of UV damage, the effect could be indirect; for example, incorporation of the CSB-PGBD3 fusion protein in place of normal CSB might facilitate disassembly of stalled TCR, DNA repair, and/or chromatin remodeling complexes, thus allowing backup repair pathways access to the DNA. In any event, the ability of the CSB-PGBD3 fusion protein to accelerate RRS after UV irradiation of CSB-null cells was so intriguing that we revisited this result using the very different HCR assay for DNA repair.

To more clearly display the effect of the fusion protein on low (and presumably more physiologically relevant) levels of UV and oxidative DNA damage, we plotted the HCR data in a new way. The log of transcriptional activity is usually plotted against a linear measure of DNA damage, although a semi-log plot exaggerates differences between cell lines at high levels of damage while minimizing differences at low levels. Instead, we normalized the HCR data to control cells expressing only the drug resistance markers, allowing the relative transcriptional activity (and thus TC-NER) to be compared over the entire range of DNA damage (Figure 2-2B,C).

More surprisingly, the CSB-PGBD3 fusion protein strongly synergized with CSB to stimulate UV repair by 200 to 250% in the HCR assay, whereas expression of the fusion protein alone had little effect and expression of the CSB protein alone rescued UV repair as expected (Figure 2-2B).

Finally, expression of CSB protein in CSB-null UVSS1KO cells rescues repair of oxidative DNA damage in the HCR assay, although expression of the CSB-PGBD3 fusion protein alone only mildly inhibits both residual oxidative repair in the absence of CSB and normal oxidative repair in the presence of CSB (Figure 2-2C). Curiously, another UV sensitive line Kps3SV13.3 that is also deficient in UV repair [38] has been shown to repair oxidative thymine glycol damage as proficiently as wild type [36].

We conclude that the CSB-PGBD3 fusion protein has significant biological activity based on its ability to stimulate UV repair in the RRS assay in the absence of CSB

(Figure 2-2A) and to synergize strongly with CSB in the HCR assay for repair of UV damage (Figure 2-2B). These DNA repair activities presumably reflect the ability of the fusion protein to influence the mechanism or pathways of DNA repair, and could help to explain why the protein has been conserved in the hominid lineage for over 43 My [14]. Alternatively, as described below, conservation could reflect the ability of the fusion protein to induce major changes in gene expression that resemble an innate antiviral immune response.

**Microarray analysis of CSB-null UVSS1KO-derived cell pools expressing CSB, CSB-PGBD3 fusion protein, both proteins, or neither**

We next examined the more general role of the fusion protein in gene expression and cell physiology using microarray analysis. RNA from UVSS1KO-derived pools expressing CSB, the CSB-PGBD3 fusion protein, both proteins, or neither (Table 2-1) was characterized using Affymetrix U133A Plus 2.0 GeneChips, and the raw data processed with the PLIER and SAM programs. As shown in Table S1A(a) and (b) — where (a) and (b) designate sheets in a workbook — genes exhibiting a robust expression change of 2-fold or more included 305 genes (388 probes) regulated by CSB, 581 genes (767 probes) regulated by CSB-PGDB3, and 1354 genes (1674 probes) regulated by CSB + CSB-PGBD3 together. The microarray data were validated by RT-PCR of selected genes (Table 2-S2) which, as is often the case, largely confirmed but occasionally diverged from the microarray values [6].

To generate an initial overview of the processes and pathways implicated by these gene expression changes, we used the L2L software suite and microarray database ([39]; www.depts.washington.edu/l2l) to examine the complete lists of robustly regulated genes (Table 2-S1A). Genes upregulated by the CSB-PGDB3 fusion protein were found to match strongly with genes upregulated by interferons (IFNs) and viral infection, more weakly with aspects of the immune response, and to a lesser extent with inflammation, apoptosis, and neural growth. In contrast, genes affected by expression of CSB alone or CSB + CSB-PGDB3 fusion protein exhibited fewer matches of comparable significance. Genes downregulated by CSB matched those regulated by IL-2 in both directions, and may reflect signaling pathways controlling proliferation (data not shown).

To carefully examine the interferon, viral, immune and other minor signatures identified by L2L analysis of all robustly regulated genes in the CSB-PGBD3 fusion protein dataset, we assembled a list of potentially relevant genes by manually interrogating the NCBI Gene Database using key phrases suggested by the L2L analysis: interferon-regulated, -induced, and -repressed; interferon α, β, γ; STAT1 and STAT; immune and

inflammatory responses; apoptosis; and neural growth and development. We then assigned each of these potentially relevant genes to 1 of the 10 functional categories indicated in columns F through O of Tables 2-S1B(a) and 2-S1B(b) based on the gene description in the NCBI Gene Database. Lastly, we culled the complete list of robustly regulated genes (Table 2-S1A) leaving only those genes that fall into at least 1 of the 10 functional categories (Table 2-S1B).

**The CSB-PGBD3 fusion protein induces a strong interferon-like response in CSB-null cells that is repressed by coexpression of CSB**

Using the binomial distribution [39] and a human gene count of 17,506 [40], we calculated the significance of overlaps between genes regulated at least 2- or 4-fold by the fusion protein (Table 2-S1A) and genes belonging to each of the 10 functional categories identified by L2L and listed in columns F through O of Tables 2-S1B(a) and 2-S1B(b). As shown in Table 2-2, expression of the CSB-PGBD3 fusion protein in CSB-null UVSS1KO cells induces a strong but atypical interferon response resembling a composite of the canonical responses to IFN-α, IFN-γ and, to a lesser extent, IFN-β. For example, of 457 genes known to be regulated, induced, or repressed by an interferon, 63 overlapped with genes induced at least 2-fold by the CSB-PGBD3 fusion protein, and 37 overlapped with genes induced at least 4-fold [41].

The list of interferon-related genes that is upregulated 2-fold or more by the CSB-PGBD3 fusion protein (Table 2-S1B(c)) includes many prominent interferon response genes: the *JAK* kinase-activated signal transducers and activators of transcription *STAT1* (7-fold) and *STAT2* (2-fold); the 2'-5'-oligoadenylate synthetases *OAS1* (118,000-fold), *OAS2* (129-fold), *OAS3* (11-fold), and *OASL* (29-fold) that activate the antviral RNase L [42]; the interferon-stimulated genes *ISG15* (11-fold) and *ISG20* (12-fold); the interferon-inducible genes *IFI6* (22-fold), *IFI27* (239-fold), *IFI44* (29-fold), *IFI44L* (338-fold), and *IFH1* (36-fold); the *IFI* genes with tetratricopeptide repeats *IFIT1* (8-fold), *IFIT2* (6-fold), *IFIT3* (4-fold), *IFIT5* (2-fold), and *IFITM1* (4-fold); and the *IRF9* subunit (6-fold) of the ISGF3 transcription factor. In addition, the IFN-α, β, and ω receptor *IFNAR2* and the receptor-activated kinase *JAK1* are induced 2.35 and 2.30-fold by coexpression of CSB and the CSB-PGBD3 fusion protein though not by either CSB nor fusion protein alone. Moreover, as might be expected from conservation of the CSB-PGBD3 fusion protein since marmosets [14], the presence of functional CSB almost completely suppresses the interferon-like response induced by fusion protein alone, reducing the p-values for overlap between fusion-induced genes and the interferon-response genes by a dramatic 21-23 orders of magnitude (Tables 2-2 and 2-S1C).

**U-STAT1 and ISGF3 appear to mediate the interferon-like response induced by the CSB-PGBD3 fusion protein in CSB-null cells**

The CSB-PGBD3 fusion protein does not appear to induce any interferon mRNAs regardless of whether the threshold of significance for the microarray data is set at a 2-fold change (Table 2-S1B) or at the 1.1-fold statistical detection limit of our biological replicates (data not shown). Thus the observed interferon-like response is unlikely to arise through the canonical JAK-STAT pathway in which interferons bind to transmembrane receptors, activating intracellular receptor-associated JAK and TYK kinases that phosphorylate STATs on tyrosine. Instead, as shown in Table 2-S1B(c), expression of the fusion protein elevates the mRNA levels for STAT1 (7-fold), STAT2 (2-fold), and IRF9 (6-fold) which together constitute the heterotrimeric transcription factor ISGF3 (interferon-stimulated gene factor 3). ISGF3 binds to ISREs (interferon-stimulated response elements) and normally drives the IFN-α and IFN-β responses, while STAT1 homodimers bind to GAS elements (IFN-γ activated sequences) and drive the IFN-γ response.

Although tyrosine phosphorylation was long thought to be essential for STAT activity, more recent work has shown that this is only true *early* in the IFN-β and IFN-γ responses [41,43,44]. Ptyr-STAT1 initially drives expression of a large number of interferon response genes, but most of these return to basal levels within 6 to 8 h presumably because continued expression would be damaging. Interestingly, Ptyr-STAT1 also induces STAT1 transcription, causing accumulation of transcriptionally active but unphosphorylated STAT1 (U-STAT1) and sustaining expression of a subset of the initial interferon-induced genes for an additional 48 to 72 h [41,44].

Remarkably, expression of the CSB-PGBD3 fusion protein in CSB-null UVSS1KO cells induces 18 of the 20 genes (Table 2-3; also see Table 2-2) most strongly induced by overexpression of the unphosphorylatable Y701F-STAT1 mutant in normal BJ fibroblasts [44]. Although the precise mechanism by which U-STAT1 sustains expression of a subset of interferon-induced genes remains to be determined, an intriguing possibility is that U-STAT1, U-STAT2, and U-IRF9 may assemble into U-ISGF3 heterotrimers analogous to phosphorylated ISGF3 formed in the initial IFN-α or IFN-β responses (H. Cheon and G.R. Stark, personal communication).

STAT phosphorylation on tyrosine (tyrosine 701 in STAT1 or the equivalent tyrosine in other STATs) was long thought to be required for subsequent phosphorylation on serine (serine 727 in STAT1 or the equivalent serine on other STATs) by one of several

kinases that can fully activate nuclear STAT1 homo- and heterodimers in response to stressors of various kinds including UV, oxidative, or other kinds of DNA damage [42,45]. However, instances are now beginning to emerge in which U-STATs play roles outside of the interferon response. For example, induction of apoptosis in cardiac myocytes by ischemia/reperfusion requires phosphorylation of STAT1 on serine 727 but not tyrosine 701 [46].

The strong correlation between genes induced by U-STAT1 in normal cells and by the CSB-PGBD3 fusion protein in CSB-null UVSS1KO cells (Table 2-3) led us to examine the phosphorylation state of STAT1 in CSB-null UVSS1KO cells stably transfected with CSB, the CSB-PGBD3 fusion protein, or neither (Table 2-1) with or without UV irradiation. As shown in Figure 2-3, nuclear and cytoplasmic fractions were probed with antibodies against STAT1, Pser-STAT1, and β-actin as a loading control. No Ptyr-STAT1α or Ptyr-STAT1β was detected with anti-Ptyr-STAT1 antibody (Figure 2-4, and data not shown), tending to rule out activation by the canonical JAK-STAT pathway, but anti-Pser-STAT1 antibody recognized Pser-STAT1α which was further induced by UV stress. Alternatively, constitutive expression of IFNs (albeit at levels below the microarray detection limit) might induce negative regulatory proteins of the SOCS (suppressor of cytokine signaling) family; the SOCS proteins could then abolish Ptyr-STAT1 by downregulating JAK kinases [44] but allow self-sustaining expression of Pser-STAT1 from the U-STAT1-responsive *STAT1* promoter [41].

**The CSB-PGBD3 fusion protein upregulates the RIG-I and MDA5 effectors of the innate, intracellular antiviral defense in CSB-null UVSS1KO cells**

The ability of the CSB-PGBD3 fusion protein to induce an interferon-like response in CSB-null UVSS1KO cells, but apparently without inducing IFN mRNAs, suggested that the fusion protein might activate an innate cytoplasmic antiviral response such as those mediated by the RIG-I (aka DDX58) and/or MDA5 proteins [47,48]. Indeed, as shown in Table 2-S1B, expression of the CSB-PGBD3 fusion protein in CSB-null UVSS1KO cells strongly upregulates *MDA5* (36-fold) and *RIG-I* (6- to 8-fold). RIG-I and MDA5 are normally activated by intracellular double-stranded or uncapped RNA indicative of viral infection, and are known to signal through the mitochondrial adaptor protein IPS-1 (IFN-β promoter stimulator 1, also called MAVS, VISA, or Cardif). Signaling can stimulate the IFN-α and/or IFN-β promoters, generating secreted interferons that activate *STAT1* through an autocrine circuit involving interferon cell-surface receptors and the canonical JAK-STAT pathway. However, RIG-I can also activate *STAT1* through a newly discovered noncanonical pathway that is independent of cell-surface interferon

receptors and possibly of the receptor-associated kinases JAK1, JAK2, and TYK2 as well [49].

Consistent with induction of an innate cytoplasmic antiviral response by the CSB-PGBD3 fusion protein in CSB-null UVSS1KO cells, one of the proteins most strongly induced by the fusion protein is BST2/tetherin. Originally known as bone marrow stromal antigen 2 (BST2), tetherin is a trans-membrane protein of the innate immune response that interferes with budding and release of enveloped viruses [50,51]. *BST2/tetherin* is induced 26,500-fold as judged by expression array analysis (Table 2-S1A) and 143-fold by the RT-PCR assay (Table 2-S2); moreover, like many other genes induced by the CSB-PGBD3 fusion protein, BST2/tetherin is repressed >200-fold by coexpression of intact CSB (Tables 2-S1A and 2-S2).

Upregulation of *RIG-I* and *MDA5* by the CSB-PGBD3 fusion protein could potentially be explained in many different ways, but one intriguing scenario would be that the CSB-PGBD3 fusion protein may deregulate CSB-dependent chromatin remodeling [6,14], thus leading to aberrant transcription, generation of double-stranded or uncapped cytoplasmic RNA, activation of RIG-I and/or MDA5, and intracellular induction of the observed atypical interferon response. Admittedly, noncanonical activation of STAT1 by RIG-I overexpression in the U937 acute myeloid leukemia (AML) cell line results in STAT1 phosphorylation on both Tyr 701 and Ser 727 [49] whereas expression of the CSB-PGBD3 fusion protein in UVSS1KO fibroblasts induces STAT1 phosphorylated on Ser 727 alone (Figures 2-4 and 2-5) as well as many of the same genes induced by the unphosphorylatable Y701F STAT1 mutant in normal BJ fibroblasts (Tables 2-2, 2-3 and 2-S3). These differences in STAT1 phosphorylation state may however be cell type-specific like so many other aspects of the interferon response [52]. The significance of the innate, intracellular antiviral response induced by the CSB-PGBD3 fusion protein in the UVSS1KO background is further supported by a direct comparison of the CS1AN and UVSS1KO microarray datasets as described below and in Tables 2-4 and 2-5.

**CSB and the CSB-PGBD3 fusion protein both induce the innate intracellular immune response to viral infection**

We had observed previously that addition of CSB to CS1AN cells (a CS patient-derived compound heterozygote expressing intact CSB-PGBD3 fusion protein and N-terminal CSB fragments) induced a chromatin remodeling signature [14]; however, the genetically equivalent addition of CSB to UVSS1KO cells expressing intact fusion protein induced an interferon-like instead of a chromatin remodeling signature. As described in Appendix A and Tables S4A and S4B, the apparent discrepancy actually

reflects database bias, i.e. under- and over-representation of particular datasets depending on the interests of those who compiled the databases, when the databases were first compiled, and how actively the databases have been curated.

To avoid the complications of microarray database bias, we compared the raw CS1AN [6] and UVSS1KO datasets *directly* to each other. We used MSigDB to convert our probe lists to gene lists in a consistent fashion, and then — because MSigDB can only compare datasets from the MSigDB database — we wrote Perl scripts to compare our datasets directly and used Excel to calculate the binomial statistics for all overlaps between genes that are regulated 2-fold or more by CSB in the compound heterozygote CS1AN, and 2-fold or more by expression of CSB, the CSB-PGBD3 fusion protein, or both proteins in CSB-null UVSS1KO cells. The control datasets were CS1AN expressing EGFP[6] and UVSS1KO expressing tags only (Table 2-1). The comparisons are shown in Table 2-5.

Three conclusions from this comparison are straightforward: CSB regulates many genes independently of the CSB-PGBD3 fusion protein — as expected if regulation of these genes requires functional CSB protein; the CSB and the CSB-PGBD3 fusion protein can work synergistically — as might be expected if the N-terminal CSB domain of the fusion protein partially mimics or modulates normal CSB functions; and (3) CSB can reverse many effects of the CSB-PGBD3 fusion protein including the induction of interferon-related genes (Table 2-S1C) — suggesting that functional CSB can displace the fusion protein from shared binding sites, and potentially explaining why the fusion protein does not behave as a dominant negative in normal individuals. Two other conclusions were unexpected, and potentially more exciting:

Expression of many genes requires coregulation by *both* CSB and the CSB-PGBD3 fusion protein — seemingly at odds with the dominance of CSB over the fusion protein as proposed above. A particularly intriguing instance of coregulation is the 7-fold induction of insulin growth factor binding protein 5 (IGFBP5) by CSB + fusion but not by CSB or fusion alone, and the concurrent 3-fold repression of IGFBP7 by CSB + fusion but not by CSB or fusion alone (Table 2-S1). IGFBPs bind insulin and related proteins, modulating or inhibiting their action. These coregulation data therefore suggest that the CSB-PGBD3 fusion protein can modulate the IGF1/insulin pathway in the presence of functional CSB, and may have been conserved not only for a role in DNA repair or chromatin remodeling [6] but for the ability to confer a metabolic advantage. These data are also consistent with induction of IGFBP1 in both aged mice and an XPF-ERCC1 progeria, and with the hypothesis that organismal resources are reallocated by the

IGF1/insulin pathway from growth to somatic preservation in response to unrepaired DNA damage [53,54].

Most intriguingly, as shown in Table 2-5, expression of CSB in the CS1AN compound heterozygote (which should restore the normal genotype) resembled expression of the CSB-PGBD3 fusion protein in the CSB-null UVSS1KO line (which should resemble the majority of CSB mutants). While trying to understand why one cell line with a nominally normal CSB genotype (CS1AN expressing CSB) would partially resemble another with a nominally mutant CSB genotype (UVSS1KO expressing CSB-PGBD3), we noticed that 15 of the 20 overlapping upregulated genes most closely matched interferon-related lists in the MSigDB database (Table 2-4A,B). In contrast, no functional themes emerged from the 15 downregulated genes (data not shown). Moreover, of the 15 upregulated genes that accounted for the overlaps, 11 belong to just 3 related functional themes — viral RNA recognition, protein degradation, and membrane-mediated antiviral activities: 4 recognize various aspects of intracellular viral RNA (*RIG-I* aka *DDX58*, *MDA5* aka *IFIH1*, *IFIT1*, *IFIT2*); 4 others are associated with protein degradation through ubiquitin-like or RING finger pathways (*HERC2*, *ISG15*, *RBBP6*, and *TRIM14* — a possible member of the *TRIM5*a, *TRIM6*, *TRIM22*, and *TRIM34* antiretroviral gene superfamily [55]; 3 others participate in membrane-related antiviral restriction (*RSAD2* aka *viperin* or *cig5* — which localizes to cytoplasmic lipid bodies and facilitates signaling through cell surface TLR7 and TLR9 nucleic acid receptors; *PLSCR1* (phospholipid scramblase) — which induces a subset of interferon-stimulated genes (ISGs) including ISG15 and guanylate binding proteins known as GBPs; and *GBP1* — a dynamin family protein involved in vesicle scission [56,57]. For a more detailed description of these innate immunity genes and functions, see Appendix 2. Taken together, these data suggest that CSB and the CSB-PGBD3 fusion protein both contribute to the cellular antiviral state and interferon-like response. Moreover, upregulation of the innate antiviral proteins RIG-I, MDA5, and BST2/tetherin by the CSB-PGBD3 fusion protein but not by CSB alone (above and Table 2-S1B) further suggests that CSB and the fusion protein have both overlapping and complementary functions in the innate antiviral immune response.

**The CSB-PGBD3 fusion protein binds MER85 elements in vitro**

Autonomous inverted terminal repeat transposons often give rise to internally deleted nonautonomous transposable elements, known as MITEs or miniature inverted terminal repeat elements, which are mobilized in trans by proteins encoded within the autonomous element [58]. Over 35 Mya, an autonomous 2.5 kb PGBD3 transposon (or a closely related piggyBac transposon) gave rise to MER85 elements [19] —

nonautonomous 140 bp elements that retain the terminal inverted repeats of the autonomous PGBD3 elements but have lost the internal transposase ORF (Figure 2-1A). These MER85s were mobilized in trans by the PGBD3 transposase and dispersed throughout the human genome in nearly 900 copies before mobility ceased ([19]; L.T. Gray, K.K. Fong, T. Pavelitz, and A.M. Weiner (2012) PLoS Genetics 8, in press).

We previously speculated that the CSB-PGBD3 fusion protein might bind MER85 elements through the C-terminal transposase domain, and that the acidic N-terminal CSB domain of the fusion protein might then influence expression of nearby genes either directly or through an effect on local chromatin structure [14]. To explore this hypothesis, we asked whether the CSB-PGBD3 fusion protein and/or solitary PGBD3 transposase are capable of binding MER85 elements in vitro.

We used the Repbase MER85 consensus (www.girinst.org/repbase/) to find all homologous elements in the March 2006 assembly of the human genome sequence (build hg18). We then PCR amplified and cloned the 6 most highly conserved MER85s (Figure 2-6, and Methods). We expressed hexahistidine-tagged CSB-PGBD3 fusion protein and PGBD3 transposase in the baculovirus system, and partially purified the proteins by cobalt chelate affinity chromatography. We assayed binding of the two proteins to the panel of MER85s by electrophoretic mobility shift assay (EMSA) as described [59]. The recombinant proteins were coincubated with end-labeled MER85 DNA fragments in the presence of nonspecific poly(dI-dC) competitor, and the resulting protein/DNA complexes resolved by native gel electrophoresis (Figure 2-7).

Although all of the MER85s conformed well to the Repbase consensus, only 4 of the 6 shifted strongly in vitro. Interestingly, the same 4 MER85s shifted with both solitary PGBD3 transposase (*middle panel*) and the CSB-PGBD3 fusion protein (*right panel*) indicating that the acidic N-terminal CSB domain did not interfere with DNA binding. Using antibodies against the N- and C-terminus of CSB, we found by ChIP-seq that all 6 MER85s bind the CSB-PGBD3 fusion protein in vivo (L.T. Gray, K.K. Fong, T. Pavelitz, and A.M. Weiner (2012), PLoS Genetics 8, in press). Thus the fusion protein may regulate gene expression both locally (by influencing gene expression near genomic binding sites) and more globally (by mimicking, modulating, or interfering with CSB functions).

## Discussion

The CSB-PGBD3 fusion protein has been conserved for >43 My from marmoset to humans, and is as highly conserved as full length CSB protein [14]. Such striking

conservation suggests that the fusion protein confers a selective advantage in the presence of functional CSB. What are these advantageous functions, what mechanisms are involved, and does the fusion protein contribute to CS disease in individuals who lack functional CSB but continue to express the fusion protein? As a first step toward answering these questions, we have examined the consequences of reintroducing the CSB-PGBD3 fusion protein, with or without functional CSB, into CSB-null UVSS1KO-derived cells which do not express either stable CSB fragments [15] or the fusion protein [6].

**The CSB-PGBD3 fusion protein is biologically active**

We have presented evidence that the fusion protein is biologically active in many respects: It modulates repair of UV and oxidative DNA damage as judged by RRS and HCR assays (Figure 2-2); it regulates expression of many genes, and coregulates additional genes together with CSB (Table 2-S1); when coexpressed with CSB, it induces insulin growth factor binding protein 5 (IGFBP5) and represses IGFBP7 (Table 2-S1), consistent with an effect on the IGF1/insulin pathway [53,54,60,61]; it induces a strong interferon-like response apparently without inducing interferon mRNAs (Table 2-2) through a U-STAT1-mediated pathway (Figures 2-4 and 2-5) that resembles the sustained response to interferon stimulation (Table 2-3); it induces an MDA5- and RIG-I-dependent innate antiviral response in the absence of RNA virus infection (Table 2-4); it binds to a large family of MER85 repetitive elements that are dispersed throughout the genome, potentially providing a mechanism for regulating expression of nearby genes (Figure 2-7); as expected from conservation of the CSB-PGBD3 fusion protein, the interferon-like and innate antiviral responses are both dramatically repressed by coexpression of intact CSB (Tables 2-2 and 2-S1C); and finally, expression of CSB in CS1AN cells that naturally express the fusion protein induces many of the same antiviral proteins [6] as expression of the fusion protein in CSB-null UVSS1KO cells (Tables 2-4A,B and 2-5), suggesting that CSB and CSB-PGBD3 fusion protein both contribute to the normal cellular antiviral state and interferon response.

**Induction of interferon-like and antiviral responses without viral infection**

How does the CSB-PGBD3 fusion protein induce interferon-like and antiviral responses without detectably inducing IFN mRNAs or activating the JAK/TYK pathway? Our current data favor three of many imaginable mechanisms: First, coinduction of all three components (STAT1, STAT2, and IRF9) of the heterotrimeric transcription factor ISGF3 (interferon-stimulated gene factor 3) suggest that the fusion protein activates a common node in the interferon response located downstream of the JAK/TYK kinases that

increases Pser727-U-STAT1 and STAT2 but not Ptyr701-STAT1 (Tables 2-S1 and 2-S3, Figures 2-4 and 2-5). Alternatively, as discussed in Results, loss of CSB chromatin-remodeling activity could lead to aberrant transcription, generation of dsRNA, and induction of the RIG-I and MDA5 innate immunity pathways that are normally induced by infection with RNA viruses or by JAK/TYK-independent pathways [49]. Whatever the mechanism(s), we speculate that the CSB-PGBD3 fusion protein may have been conserved for 43 My because it is able to prime or poise the interferon and/or innate immune or antiviral responses in which speed may be critical for success. Finally, we cannot rule out the scenario (also discussed in Results) in which very low levels of constitutive IFN expression induce SOCS-family proteins, downregulating JAK kinases [44] but allowing expression of U-STAT1 from the *STAT1* promoter [41].

The phosphorylation of STAT1 on serine 727 without prior phosphorylation of tyrosine 701 (Figure 2-4 and Figure 2-5) is unusual but not unprecedented. STAT phosphorylation is known to be regulated by several serine kinases including ERK (extracellular signal-regulated protein kinase), p38, JNK (JUN N-terminal kinase), PKCδ (protein kinase Cδ), and possibly CAMK2 (calcium/calmodulin dependent kinase II); and JAK-STAT signalling can be regulated by a variety of cellular signaling pathways through SOCS proteins, PIAS family proteins (protein inhibitors of activated STAT), and various PTPs (protein tyrosine phosphatases) [62,63]. Perhaps most surprisingly, the innate immune response to cytoplasmic dsRNA is severely attenuated in human embryonic stem cells because certain key proteins are absent and others cannot be activated [64]. The developmental and tissue specificity of STAT activation, as well as the diversity of signaling inputs, are almost certain to increase the variety of regulatory nodes downstream of JAK/TYK kinases by which CSB-PGBD3 expression could activate interferon-like and antiviral responses in the absence of interferons and viral infection.

**Possible relevance of the CSB-PGBD3 fusion protein to CS disease**

Many nonsense and frameshift mutations within the N-terminal CSB domain of the CSB-PGBD3 fusion protein are known that prevent synthesis of the fusion protein, yet still cause CS; and there does not appear to be any correlation between continued expression of the CSB-PGBD3 fusion protein [2] and the surprisingly heterogeneous clinical presentation of CS patients [34]. Nevertheless, the ability of the fusion protein to modulate DNA repair and to induce an interferon-like innate antiviral response in UVSS1KO cells suggest that the fusion protein could contribute to CS especially in patients from consanguineous backgrounds where (as discussed in the Introduction) partial homozygosity uncovers some of the most divergent CS phenotypes [15,32,33].

Alternatively, Brooks et al. [65] have noted that several neurodegenerative diseases including Trichothiodystrophy (TTD), Aicardi-Goutières syndrome (AGS), and CS exhibit characteristic dysmyelination, calcification, and microcephaly. In TTD, causative mutations in the XPD component of TFIIH reduce TFIIH coactivator function on myelin-related genes. In AGS, mutations in the *TREX1* or *RNASEH2* nucleases cause accumulation of S-phase DNA fragments that induce a type I interferon response through the STING-dependent innate antiviral response [66]; and we show here that expression of the CSB-PGBD3 fusion protein in CSB-null UVSS1KO cells, or expression of intact CSB in patient-derived CS1AN cells that naturally express the fusion protein, both induce a similar cohort of innate antiviral genes (Table 2-4 and Appendix 2). As all 5 genes that cause CS (*CSA*, *CSB*, *XPB*, *XPD*, and *XPG*) are subunits of TFIIH [10], and expression of the CSB-PGBD3 fusion protein or intact CSB can induce an interferon-like antiviral response in certain genetic backgrounds, it is possible that transcriptional dysregulation and/or an inappropriate interferon response may contribute to the remarkable heterogeneity in CS onset and symptoms.

**The CSB-PGBD3 fusion protein may confer a metabolic advantage**

Niedernhofer et al. [54] and van der Pluijm et al. [53] have suggested that CS may reflect reallocation of resources from growth to somatic preservation by the IGF1/insulin pathway in response to unrepaired DNA damage. Our data may support this hypothesis. We find that CSB and the CSB-PGBD3 fusion protein together, but neither protein alone, induce IGFBP5 and repress IGFBP7 (Table 2-S1A) implying that the fusion protein can modulate the IGF1/insulin pathway in normal cells which have functional CSB. Similarly, IGFBP1 is induced in aged mice and an XPF-ERCC1 progeria [53,54]; and the Drosophila IGFBP7 homolog binds insulin-like peptides in vivo, downregulates insulin/IGF signaling, and prolongs lifespan [67]. The discovery that CSB and CSA are required for mitochondrial as well as nuclear base excision repair (BER) potentially explains why CSB mutations increase cellular reactive oxygen species (ROS) and suggests that the effects of defective nuclear TCR, and activation of the IGF1/insulin pathway, may be compounded by mitochondrial dysfunction [68]. In any event, modulation of the IGF1/insulin pathway by the CSB-PGBD3 fusion protein in the presence of functional CSB may suggest that the conserved fusion protein confers a metabolic advantage rather than, or in addition to, effects on DNA repair and/or chromatin remodeling.

**A cautionary note regarding the emergent functions of other human fusion proteins**

Although the CSB-PGBD3 fusion protein has been conserved for >43 My, shares the same N-terminal 465 residues as CSB, and is coexpressed with CSB by alternative splicing [6,14], the selective advantage of the fusion protein need not be related to CSB function in normal cells. Thus, although our data suggest that the fusion protein may compete with or modulate CSB functions in normal cells and/or affect CSB-related functions in individuals lacking functional CSB protein, the fusion protein could also have *emergent* functions that differ from the normal functions of the component proteins.

Consider the two other human fusion proteins that have been studied in some detail: (1) The NUP98-HOXA9 fusion protein, which is generated by a chromosomal rearrangement joining the N-terminal FG-repeat domain of nucleoporin NUP98 to the C-terminal DNA-binding domain of the HOXA9 homeodomain transcription factor, causes acute myeloid leukemia (AML) [69]. Although the fusion protein can function as a transcriptional activator targeted by the C-terminal DNA-binding HOXA9 domain [70], this is *not* the cause of AML. Rather, the N-terminal NUP98 FG-repeat domain of NUP98-HOXA9 forms intranuclear aggregates that sequester the exportin CRM1 [71] resulting in constitutive expression of transcription factors such as NFAT and NFκB that are normally downregulated by nuclear export. (2) The SETMAR fusion protein (aka Metnase) emerged 40–58 Mya and is generated by splicing of a functional histone methyltransferase (SET) domain to a mariner transposase (MAR) which retains specific binding to mariner terminal inverted repeats (TIRs) in vitro [72]. Strikingly, SETMAR/Metnase tethers a chromatin-altering histone methylase (SET) domain to dispersed mariner inverted repeats, just as the CSB-PGBD3 fusion protein tethers the potentially chromatin-altering acidic N-terminal domain of CSB to dispersed MER85 repeats. Yet the only known activity of SETMAR/Metnase is to facilitate nonhomologous end joining (NHEJ), a global repair function that requires both the histone methyltransferase of the SET domain [73] and an endonuclease activity of the mariner transposase but *not* the capacity for site-specific DNA binding [74]. Thus the ability of SETMAR/Metnase to facilitate NHEJ, and NUP98-HOXA9 to cause AML, both reflect *emergent* functions — and the same could be true for the CSB-PGBD3 fusion protein.

## Materials and methods

### DNA constructs

The parent for all expression constructs was the bicistronic pIREShyg3 vector (Clontech). A N-terminal 3 x FLAG tag was inserted to generate pFLAG-IREShyg, followed by an HA tag to generate pFLAG-HA-IREShyg. Open reading frames for intact

CSB (4.5 kb) and the CSB-PGBD3 fusion protein (3.2 kb) were inserted downstream of the tags to generate pFLAG-HA-CSB-IREShyg and pFLAG-HA-CSB-PGBD3-IREShyg. To generate pFLAG-CSB-PGBD3-IRESneo, the $hyg^R$ gene of pFLAG-HA-IREShyg was replaced by the $neo^R$ gene from pIRESneo3, and the open reading frame of the CSB-PGBD3 fusion protein was inserted downstream of the tags. Details are available upon request.

**Stably transfected pools**

pFLAG-HA-CSB-IREShyg3, pFLAG-HA-CSB-PGBD3-IREShyg3, and the empty vector pFLAG-HA-IREShyg3 were linearized with XhoI just downstream from the poly(A) site, and transfected using TransIT-LT1 reagent (Mirus) into UVSS1KO cells grown in DMEM. Selection with 100, 150 and 200 μg/ml hygromycin (Invitrogen) was begun after 24 h, and both drug and media were refreshed every 48-72 h. Confluent wells containing 50-100 colonies were trypsinized and passaged thereafter as pools. To generate doubly-transfected cells for the HCR experiments, the singly-transfected hygromycin-resistant pools were transfected with either pFLAG-CSB-PGBD3-neo linearized by PvuI within the $amp^R$ gene, or with pFLAG-neo linearized with SpeI just upstream from the CMV promoter. G418 selection was increased from 200 to 600 μg/ml while continuing 200 μg/ml hygromycin selection. We were unable to obtain hyg + CSB-neo or fusion-hyg + CSB-neo cells that expressed readily detectable CSB protein. For clarity, the cell lines used in this and the previous study [6] are listed in Table 2-1.

**Western blots**

To assay CSB and CSB-PGBD3 expression (Figure 2-3), subconfluent cells were harvested, resuspended in 2  SDS sample loading buffer, and immediately heated to 100°C for 10 min. To assay STAT1 expression, cytoplasmic and nuclear fractions were prepared as described [75,76]. To examine the UV response, adherent cells were washed in PBS, subjected to 40 $J/m^2$ UV irradiation, and allowed to grow for 30 min in fresh medium before harvest [77]. SDS-PAGE and Western blots were as described previously [14]. CSB and the CSB-PGBD3 fusion protein were detected with an antigen-purified rabbit polyclonal raised against CSB residues 1-240 [78]. The β-actin loading and dilution control was detected with a mouse monoclonal antibody (Sigma-Aldrich A2228). STAT1, phospho-STAT1(Tyr701) and phospho-STAT1(Ser727) antibodies were used to identify STAT1 phosphorylation states (Cell Signaling Technology #9172, #9171 and #9177). HRP-conjugated secondary antibodies were goat anti-rabbit and anti-mouse (ThermoScientific #31460 and #31430).

**Recovery of RNA synthesis (RRS) assays**

Cells were grown in 24 well microtiter plates under 200 μg/ml hygromycin selection before irradiation and during recovery. The cells were washed in PBS, subjected to 10 $J/m^2$ UV irradiation under a germicidal lamp, and immediately immersed in 1 ml unlabeled medium. For recovery times of 2, 6, 12 and 24 h as well as for unirradiated controls, unlabeled medium was replaced with DMEM containing 10 μCi/ml of [5,6-$^3$H]-uridine (GE Healthcare) followed by pulse-labeling for 1 h at 37ºC. Samples were processed as described [15]. Scintillation data were normalized to cell number, plotted, and standard errors of the mean calculated using Excel and GraphPad Prism. All assays were performed in triplicate. UV irradiation was calibrated using an Ultraviolet Meter (UVP).

**Host cell reactivation (HCR) assays**

UV- and $OsO_4$-damaged pEGFP-IRESpuro plasmid templates were prepared as described [36]. Transfections using Fugene 6 (Roche) were performed in quadruplicate for each level of DNA damage. EGFP fluorescence was measured using a plate reader and the values corrected for cell number based on protein content as determined by the BCA Protein Assay Kit (Pierce). The corrected fluorescence measurements were averaged and normalized to the corresponding untreated controls. The data from two independent experiments were averaged, plotted, and standard errors of the mean calculated using Excel and GraphPad Prism.

**Expression array protocol and data analysis**

Sample preparation for expression array analysis, data generation by the Center for Expression Arrays (University of Washington), and RT-PCR validation have been described previously [6]. Three independent preparations of total RNA from the CSB, CSB-PGBD3, CSB + CSB-PGBD3, and tag-only cells (Table 2-1) were quality-controlled, labeled, and used to probe Affymetrix GeneChip® Human Genome U133 Plus 2.0 Arrays. The 12 datasets were normalized using the Probe Logarithmic Intensity Error (PLIER) program of the Affymetrix Expression Console v1.1.1. Using the Significance Analysis for Microarrays (SAM) program of the MeV v4.4 software suite (www.TM4.org), fold changes for expression of individual probes were calculated by comparing the 3 datasets for one pool to the 3 datasets for another, and the resulting 9 pairwise fold changes were averaged to give the fold change for that probe.

**Mobility shift assays**

Open reading frames encoding the PGBD3 transposase and CSB-PGBD3 fusion protein [6] were cloned into the pFastBAC HT baculovirus vector (Invitrogen Bac-to-Bac® Baculovirus Expression System). Virus production and protein expression in SF9 cells were performed as recommended by the supplier. Soluble hexahistidine tagged

protein was partially purified over a TALON® resin (Clontech), eluted with imidazole HCl, desalted, and concentrated by Centricon filtration. Six different MER85s that closely matched the 140 bp Repbase consensus (www.girinst.org/repbase/) were amplified by genomic PCR using an upstream primer with a BamHI site and a downstream primer with EcoRI; the upstream flank varied from 82-189 bp, the downstream flank from 187-326 bp (Figure 2-6). The PCR fragments were cloned between the BamHI and EcoRI sites of pBluescript, excised by restriction digestion, and [$^{32}$P]-labeled by filling in the ends. Mobility shift assays were performed as described [59].

# Figures

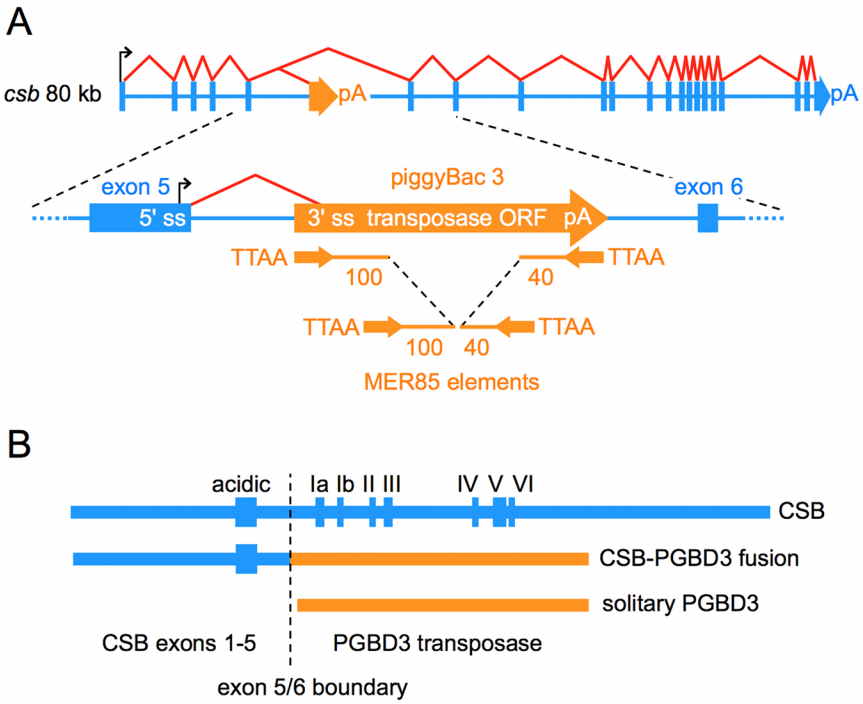



**Figure 2-1. piggyBacs are mobile DNA elements that survive as alternative 3' exons.**
(A) PGBD3 inserted into intron 5 of the primate *CSB* gene at least 43 Mya in the common ancestor of simian primates, with the result that the *CSB* gene now generates three proteins as shown in (B): full length CSB by default splicing of all 22 CSB exons, CSB-PGBD3 fusion protein by alternative splicing between CSB exon 5 and the PGBD3 alternative 3' terminal exon, and solitary PGBD3 driven by a cryptic promoter in CSB exon 5 [14]. The PGBD3 insertion generated a TTAA target site duplication. Immediately inside the subterminal inverted repeats of the mobile element, the transposase open reading frame (ORF) is flanked upstream by a 3' splice site (3' ss) and downstream by a polyadenylation site (pA). MER85 elements are nonautonomous internally-deleted PGBD3-derived elements that were last mobilized by a PGBD3-like transposase about 35 Mya [19]. CSB and PGBD3 sequences are indicated in blue and orange, respectively. The schematic not drawn to scale; the CSB gene spans 80 kb, the PGBD3 element 2.5 kb, and intact MER85s only 140 bp. (B) A comparison of the three proteins encoded by the CSB locus. The fusion protein joins the acidic 465 N-terminal residues of CSB exons 1-5, but none of the ATPase motifs (Roman numerals), to the 595 residue PGBD3 transposase.

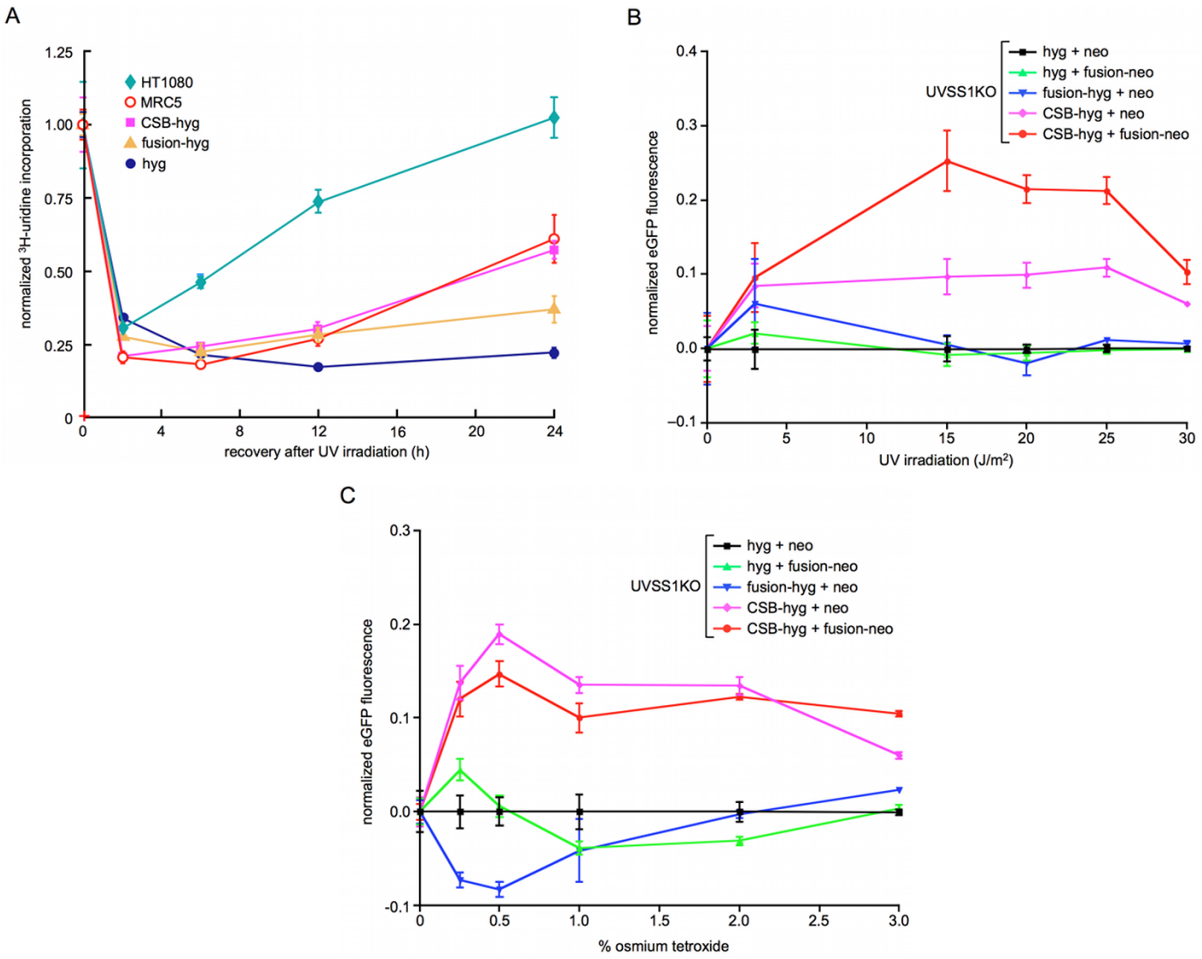



**Figure 2-2. Recovery of RNA synthesis (RRS) following UV damage, and host cell reactivation (HCR) assays for repair of oxidation- and UV-induced DNA damage.** Assays were performed using CSB-null UVSS1KO-derived pools stably expressing CSB, CSB-PGBD3 fusion protein (two independent pools selected with either hygromycin or neomycin), both proteins, or tags only (Table 2-1). (A) RRS assays monitoring $^{3}$H-uridine incorporation after UV irradiation of growing cells with 10 J/m$^2$/min. The SV40-immortalized normal lung cell line MRC5-SV and the the HT1080 fibrosarcoma cell line were included as a controls [79]. (B) HCR assays for UV damage. The EGFP expression construct was irradiated with 1 J/m$^2$/min UV from a germicidal lamp for 5-30 min before transfection. EGFP fluorescence was measured 48 h after transfection and normalized to cell number at each time point to correct for cell growth. (C) HCR assays for oxidative damage. The reporter EGFP expression construct was pretreated with the indicated concentrations of osmium tetroxide before transfection. Assays were performed in triplicate (panel A) or octuplicate (panels B and C); error bars smaller than the datapoint icons are not shown.

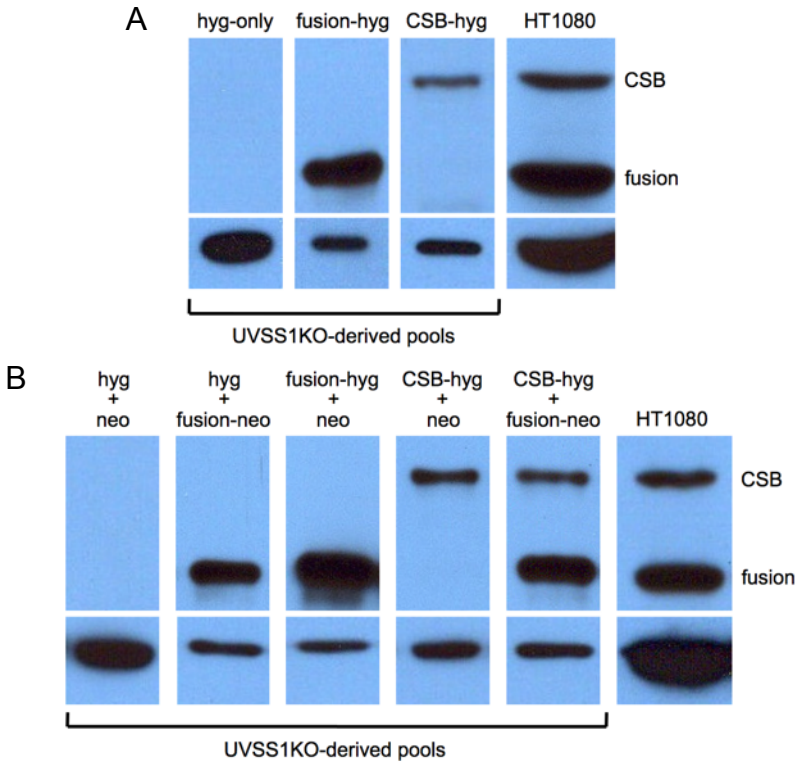



**Figure 2-3. Overexpression of CSB and CSB-PGBD3 fusion proteins in the stably transfected UVSS1KO-derived pools.**
To avoid protein degradation, whole cell lysates were prepared from each pool by plunging harvested cells into hot SDS (see Materials and methods). An antigen-purified, rabbit polyclonal antibody directed against the N-terminus of CSB [14] was used to insure comparable signals in the Western blots for full length CSB and the CSB-PGBD3 fusion protein. The proteins were 4- to 8-fold more highly expressed in the stable UVSS1KO-derived cells than in the human HT1080 fibrosarcoma as judged by the appropriate dilution series; only representative panels from those blots are shown here. (A) Singly transfected pools used in the RRS assay (Figure 2-2A). The CSB and CSB-PGBD3 fusion proteins are N-terminally tagged with FLAG-HA, but detected here with N-terminal CSB antibody to allow comparison with endogenous expression. The HT1080 lane was overloaded so that endogenous CSB and CSB-PGBD3 signals would be comparable to what is seen in the UVSS1KO-derived cells. The hyg-only lane was also overloaded to demonstrate the absence of CSB and CSB-PGBD3 in the parental CSB-null UVSS1KO cells. (B) Doubly transfected pools used in HCR assays (Figure 2-2B and C). As described in Materials and methods, these pools were generated by transfecting first with FLAG-HA-tagged proteins or the FLAG-HA tags alone under hygromycin selection, then with FLAG-tagged proteins or the FLAG tag alone under neomycin and continued hygromycin selection. The HT1080 and hyg + neo lanes are overloaded as in (A).

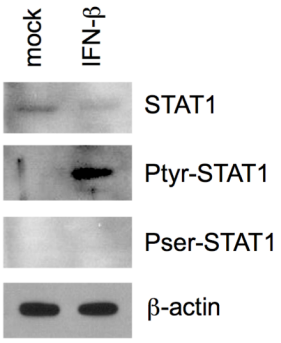



**Figure 2-4. Anti-Ptyr-STAT1 antibody is functional and can detect Tyr701-phosphorylated STAT1 early after IFN-β induction.**
The adherent human 2fTGH fibroblast line was treated with 100 IU/ml of human interferon-β (IFN-β) or vehicle alone (Mock) for 16h. Cell lysates were resolved by SDS-PAGE, electroblotted onto a PVDF membrane, and probed with the indicated antibodies. As previously observed [80], Ptyr-STAT1 accumulates before Pser-STAT1. β-actin served as a loading control.

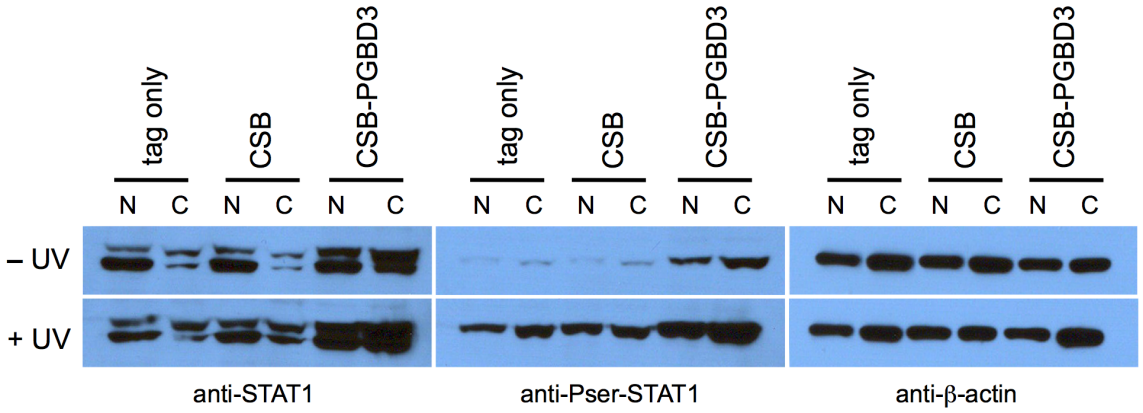



**Figure 2-5. Expression of the CSB-PGBD3 fusion protein in CSB-null UVSS1KO cells induces U-STAT1.**
USTAT1 can be phosphorylated on serine 727 and further induced by UV irradiation. Using CSB-null UVSS1KO cells stably transfected with CSB, CSB-PGBD3 fusion protein, or tags alone (Table 2-1), nuclear and cytoplasmic fractions were resolved by SDS-PAGE, blotted, and probed with anti-STAT1 (*left panel*), anti-Pser-STAT1 (*middle panel*), and anti-β-actin antibodies as a loading control (*right panel*). The STAT1 locus generates two proteins (*left panel*): full length STAT1α (*upper band*) and C-terminally deleted STAT1β (*lower band*) lacking the Ser727 phosphorylation site. STAT1β has been referred to as a splice variant or proteolysis product [81]; however, mRNAs annotated on the UCSC Genome Browser (build hg18) indicate that STAT1β reflects alternative polyadenylation within intron 23. This results in use of a TAA terminator immediately following exon 23, and a protein that retains Tyr701 but lacks the Ser727 phosphorylation site.

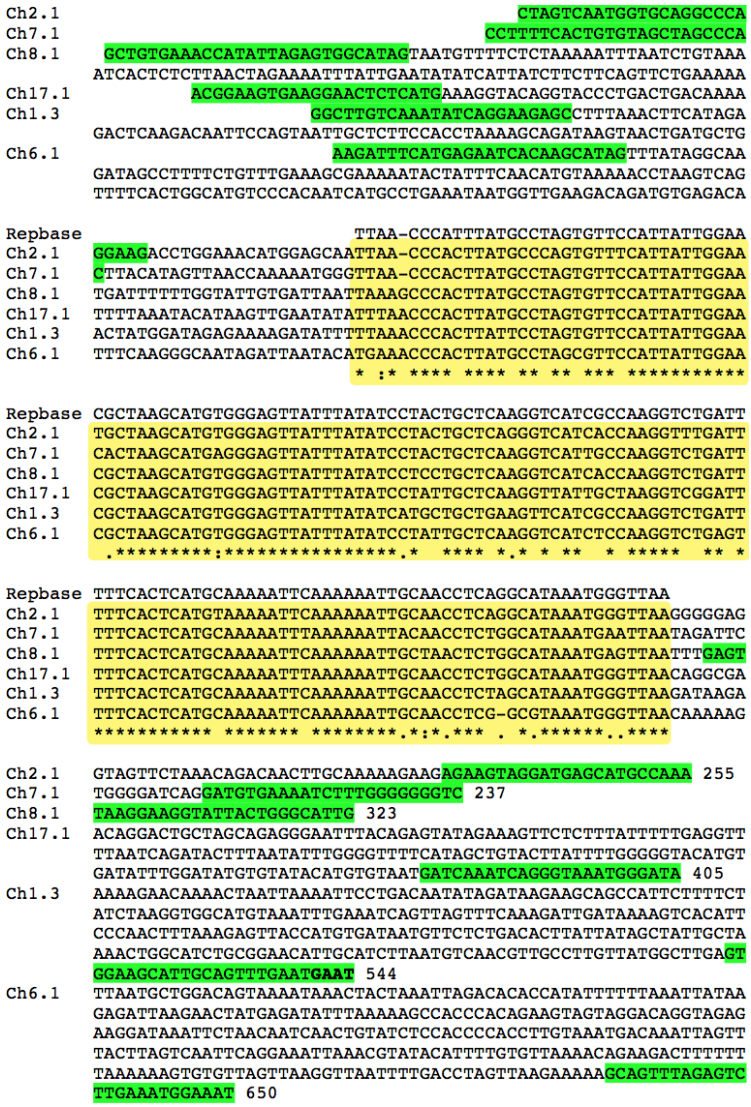



**Figure 2-6. Clustal 2.1 multiple sequence alignment of 6 representative MER85 elements with a good match to the Repbase consensus.**

The nomenclature "Ch1.3" indicates the MER85 on chromosome 1 with the third best match to the Repbase consensus (www.girinst.org/repbase/) when both length and sequence are taken into account. Yellow highlighting indicates genomic MER85 sequences including the terminal TTAA target site duplication and the subterminal inverted repeats (CCCATTTATGCCT... AGGCATAAATGGG). *, sequence match; :, identical base pair; ., maintains register but not base pair. Only genomic sequences are shown; the PCR primers also introduced 5' BamHI and 3' EcoRI sites. The last 4 nucleotides of the Ch1.3 sequence (shown in bold) contribute the first 4 nucleotides of the 3' EcoRI site. The cloned fragment lengths (shown at the 3' end of each sequence) do not include the terminal nongenomic restriction sites provided by the primers. The 5' flanking genomic sequences are collected in single panel above, and 3' flanking genomic sequences in a single panel below.

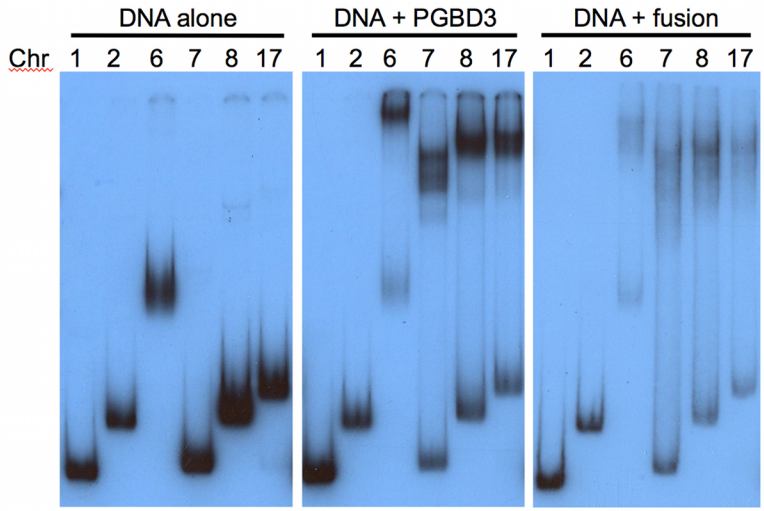



**Figure 2-7. The PGBD3 piggyBac transposase and the CSB-PGBD3 fusion protein bind to consensus MER85 elements *in vitro*.**
Electrophoretic mobility shift assays were performed [59] for binding of the recombinant PGBD3 and CSB-PGBD3 fusion proteins to 6 different genomic MER85 elements that closely match the 140 bp Repbase MER85 consensus. The multiple sequence alignment and genomic primers are shown in Figure 2-S2. Chromosome number is indicated above the lanes. MER85 elements from chromosomes 1 and 7 are MseI/MseI and AseI/HincII fragments, respectively, lacking flanking sequence; MER85s from chromosomes 2, 6, 8, and 17 are BamHI/EcoRI fragments that include the flanks. Although both proteins shift efficiently, the piggyBac transposase generates sharper bandshifts than the larger fusion protein on low percentage gels (6% 29:1 acrylamide:bisacrylamide for PGBD3 and 5% 80:1 acrylamide:bisacrylamide for CSB-PGBD3).

| Cell line | Source | Use in this study |
|---|---|---|
| CS1AN-hTERT | {Newman:2006gd} | |
| CS1AN-hTERT + eGFP-puro | {Newman:2006gd} | Table 2-4 |
| CS1AN-hTERT + CSB-puro | {Newman:2006gd} | Table 2-4 |
| UVSS1KO-SV40 + hyg | this study | Figure 2-2A, Tables 2-4 and 2-S1 |
| UVSS1KO-SV40 + fusion-hyg | this study | Figure 2-2A, Tables 2-4 and 2-S1 |
| UVSS1KO-SV40 + CSB-hyg | this study | Figure 2-2A, Tables 2-4 and 2-S1 |
| UVSS1KO-SV40 + hyg + neo | this study | Figure 2-2B,C |
| UVSS1KO-SV40 + hyg + fusion-neo | this study | Figure 2-2B,C |
| UVSS1KO-SV40 + fusion-hyg + neo | this study | Figure 2-2B,C |
| UVSS1KO-SV40 + CSB-hyg + neo | this study | Figure 2-2B,C |
| UVSS1KO-SV40 + CSB-hyg + fusion-neo | this study | Figure 2-2B,C, Tables 2-4 and 2-S1 |

**Table 2-1. Cell lines used in this and the previous study.**
The primary GM00739 fibroblast line derived from patient CS1AN was obtained from the Coriell Institute and transformed with retroviral hTERT [6]. UVSS1KO fibroblasts derived from patient UVSS1KO and transformed with a replication-defective SV40 [15] were the kind gift of Kiyoji Tanaka (Osaka University). EGFP, enhanced green fluorescent protein; hyg, hygromycin resistance; neo, neomycin resistance; fusion, CSB-PGBD3 fusion protein. Puromycin-resistant cells express untagged proteins from the bicistronic vector pIRESpuro (Clontech); hygromycin-resistant cells express FLAG-HA tagged proteins or the FLAG-HA tag alone; neomycin-resistant cells express FLAG-tagged proteins or the FLAG tag alone. Hyphens indicate transformation protocol for cell lines, antibiotic selection for transfection of genes and tags. The microarray experiments (Table 2-S1) were performed using singly-transfected CSB-hyg and fusion-hyg pools, and a doubly-transfected CSB-hyg + fusion-neo line, all normalized for consistency to hyg alone. See Materials and methods for details.

| GO terms and processes | | Genes in GO list | CSB | | CSB + fusion | | fusion | |
|---|---|---|---|---|---|---|---|---|
| | fold induction | | ≥ 2 fold | ≥ 4 fold | ≥ 2 fold | ≥ 4 fold | ≥ 2 fold | ≥ 4 fold |
| | genes affected | | 305 | 93 | 1354 | 219 | 581 | 143 |
| IFN regulated/induced/repressed | | 457 | 10 | 2 | 45 | 7 | 63 | 37 |
| | | | 2.78E-01 | 6.98E-01 | 6.34E-02 | 3.48E-01 | 5.49E-21 | 8.36E-25 |
| IFN-a regulated/induced/repressed | | 74 | 1 | 1 | 10 | 1 | 22 | 16 |
| | | | 7.25E-01 | 3.26E-01 | 6.58E-02 | 6.05E-01 | 2.39E-14 | 3.83E-18 |
| IFN-b regulated/induced/repressed | | 40 | 1 | 0 | 3 | 1 | 10 | 6 |
| | | | 5.02E-01 | 1.00E+00 | 3.74E-01 | 3.94E-01 | 1.33E-06 | 1.16E-06 |
| IFN-g regulated/induced/repressed | | 242 | 6 | 2 | 22 | 6 | 29 | 18 |
| | | | 2.49E-01 | 3.68E-01 | 2.52E-01 | 8.54E-02 | 5.96E-09 | 3.06E-12 |
| IFN-e regulated/induced/repressed | | 53 | 1 | 1 | 5 | 2 | 7 | 2 |
| | | | 6.03E-01 | 2.46E-01 | 3.91E-01 | 1.43E-01 | 2.22E-03 | 7.04E-02 |
| IFN-k regulated/induced/repressed | | 97 | 2 | 0 | 7 | 1 | 10 | 8 |
| | | | 5.04E-01 | 1.00E+00 | 6.23E-01 | 7.04E-01 | 1.77E-03 | 1.63E-06 |
| IFN-w regulated/induced/repressed | | 20 | 0 | 0 | 1 | 0 | 3 | 1 |
| | | | 1.00E+00 | 1.00E+00 | 7.87E-01 | 1.00E+00 | 2.98E-02 | 1.51E-01 |
| STAT1 regulated/induced/repressed | | 192 | 7 | 1 | 27 | 5 | 18 | 12 |
| | | | 5.30E-02 | 6.40E-01 | 2.75E-03 | 9.50E-02 | 1.06E-04 | 8.39E-08 |
| STAT regulated/induced/repressed[a] | | 257 | 7 | 1 | 26 | 4 | 22 | 9 |
| | | | 1.65E-01 | 7.46E-01 | 1.05E-01 | 4.01E-01 | 7.15E-05 | 3.10E-04 |
| U-STAT1 prolonged expression[b] | | 108 | 5 | 2 | 16 | 3 | 49 | 29 |
| | | | 4.21E-02 | 1.13E-01 | 1.17E-02 | 1.54E-01 | 1.20E-38 | 7.16E-35 |
| STAT1 prolonged expression[b] | | 35 | 2 | 0 | 4 | 0 | 23 | 19 |
| | | | 1.25E-01 | 1.00E+00 | 2.87E-01 | 1.00E+00 | 2.68E-22 | 8.66E-29 |
| Immune response | | 1330 | 35 | 5 | 99 | 11 | 79 | 37 |
| | | | 1.23E-02 | 8.34E-01 | 6.69E-01 | 9.43E-01 | 9.00E-07 | 3.23E-10 |
| Immune response[c] | | 1043 | 27 | 4 | 73 | 7 | 33 | 6 |
| | | | 3.00E-02 | 8.04E-01 | 8.25E-01 | 9.75E-01 | 6.34E-01 | 8.53E-01 |
| Inflammatory response | | 1190 | 82 | 16 | 21 | 8 | 75 | 29 |
| | | | 8.74E-01 | 4.20E-01 | 5.06E-01 | 3.01E-01 | 1.84E-07 | 3.76E-07 |
| Regulation of apoptosis | | 2129 | 48 | 18 | 165 | 32 | 96 | 29 |
| | | | 4.66E-02 | 3.98E-02 | 5.01E-01 | 1.70E-01 | 2.02E-03 | 6.48E-03 |
| Induction of apoptosis | | 1054 | 22 | 8 | 80 | 15 | 69 | 23 |
| | | | 2.25E-01 | 2.02E-01 | 5.85E-01 | 3.43E-01 | 1.40E-07 | 3.16E-05 |
| Neural development | | 432 | 13 | 5 | 38 | 8 | 24 | 7 |
| | | | 4.22E-02 | 8.26E-02 | 2.33E-01 | 1.78E-01 | 1.11E-02 | 6.67E-02 |
| Neural growth | | 270 | 13 | 5 | 29 | 8 | 26 | 7 |
| | | | 1.07E-03 | 1.54E-02 | 5.20E-02 | 2.15E-02 | 2.13E-06 | 7.29E-03 |

[a]includes all STATs

[b]from Table 1 of Cheon et al. [31] U-STAT1 also known as YF-STAT1

[c]excludes IFN regulated/induced/repressed genes

**Table 2-2. Expression of the CSB-PGBD3 fusion protein in CSB-null UVSS1KO cells induces an atypical interferon-like response.**

The binomial probability of overlap is shown between genes robustly regulated by the fusion protein (Table 2-S1A) and genes belonging to each of the 10 functional categories indicated in columns F through O of Tables 2-S1B(a) and 2-S1B(b). The most significant p-values (p<6E-09) are highlighted in orange; somewhat less significant p-values (3E-04>p>7E-08) are highlighted in yellow.

| Gene | Gene Description | CSB + fusion | CSB | fusion |
|---|---|---|---|---|
| | | | | |
| IFI27 | IFNa-inducible protein 27 | --- | --- | 238.83 |
| BST2 | bone marrow stromal cell antigen 2 | --- | --- | 26483.17 |
| OAS1 | 2',5'-oligoadenylate synthetase 1, 40/46kDa | --- | --- | 118155.11 |
| OAS2 | 2'-5'-oligoadenylate synthetase 2, 69/71kDa | --- | --- | 129.38 |
| OAS3 | 2'-5'-oligoadenylate synthetase 3, 100kDa | --- | --- | 11.03 |
| STAT1 | signal transducer and activator of transcription 1, 91kDa | 2.66 | --- | 6.92 |
| IFI44 | IFN-induced protein 44 | --- | --- | 29.45 |
| IFI44L | IFN-induced protein 44-like | --- | --- | 337.68 |
| IFIH1 | IFN induced with helicase C domain 1 | --- | --- | 35.60 |
| IFITM1 | IFN induced transmembrane protein 1 (9-27) | --- | --- | 4.34 |
| IFI35 | IFN-induced protein 35 | --- | 0.45 | 3.40 |
| IFIT3 | IFN-induced protein with tetratricopeptide repeats 3 | 2.36 | --- | 4.01 |
| MX1[a] | myxovirus (influenza virus) resistance 1 | --- | --- | 82.59 |
| IRF7 | IFN regulatory factor 7 | --- | 0.50 | --- |
| ISG15[b] | ISG15 ubiquitin-like modifier | 0.39 | --- | 11.32 |
| IFIT1 | IFN-induced protein with tetratricopeptide repeats 1 | --- | --- | 8.35 |
| PLSCR1 | phospholipid scramblase 1 | 2.12 | --- | 6.23 |
| HERC6 | hect domain and RLD 6 | --- | --- | 4.30 |
| FLJ20035[c] | hypothetical protein FLJ20035 | --- | --- | 3.67 |
| EPSTI1 | epithelial stromal interaction 1 (breast) | --- | --- | --- |
| | | | | |
| [a]also known as IFN-inducible protein p78 (mouse) | | | | |
| [b]also known as G1P2 | | | | |
| [c]also known as DDX60 | | | | |

**Table 2-3. Overlap between genes induced by the CSB-PGBD3 fusion protein in CSB-null UVSS1KO cells and genes induced by U-STAT1 in normal cells.**
Y701F-STAT1 (also known as U-STAT1) lacks Y701 and can only be phosphorylated on S727. Genes induced by overexpression of U-STAT1 in STAT1 normal cells are taken from Cheon et al. [44]; genes induced by CSB-PGBD3 fusion protein in CSB-null UVSS1KO background are from Table 2-S1A.

**A. MSigDB IFN-Related Datasets**

| Gene | MOSERLE[a] | DAUER[b] | BROWNE[c] | BROWNE[d] | DER[e] | DER[f] | DER[g] |
|---|---|---|---|---|---|---|---|
| DDX58 | | UP | DN | | | | |
| GBP1 | UP | | | UP | UP | UP | UP |
| HERC5 | UP | DN | | | | | |
| IFIH1 | UP | DN | | | | | |
| IFIT1 | UP | DN | UP | | | | |
| IFIT2 | UP | | UP | UP | | | |
| IFITM1 | UP | | | | UP | UP | UP |
| ISG15 | | DN | UP | UP | UP | UP | UP |
| PHLDA1 | | | | | UP | | UP |
| PLSCR1 | | | | UP | | | |
| PMAIP1 | | | UP | | UP | UP | UP |
| RBBP6 | | | UP | | | | |
| RSAD2 | UP | | UP | UP | | | |
| SAMD9 | UP | DN | | | | | |
| TRIM14 | | DN | | UP | UP | UP | |

[a]MOSERLE_IFNA_RESPONSE [46]
[b]DAUER_STAT3_TARGETS_DN [47]. Note that STAT3 represses IFN-inducible genes involved in wound healing and cancer.
[c]BROWNE_HCMV_INFECTION_4HR_UP [48]
[d]BROWNE_INTERFERON_RESPONSIVE_GENES [48]
[e]DER_IFN_BETA_RESPONSE_UP [49]
[f]DER_IFN_ALPHA_RESPONSE_UP [49]
[g]DER_IFN_GAMMA_RESPONSE_UP [49]

**B. UVSS1KO datasets**

| Gene | fusion[a] | CSB | CSB + fusion |
|---|---|---|---|
| DDX58 | UP[b] | | |
| GBP1 | UP | UP | UP |
| HERC5 | | | |
| IFIH1 | UP | | |
| IFIT1 | UP | | |
| IFIT2 | UP | | |
| IFITM1 | UP | | |
| ISG15 | UP | | |
| PHLDA1 | UP | | |
| PLSCR1 | UP | UP | |
| PMAIP1 | UP | UP | |
| RBBP6 | | | |
| RSAD2 | UP | | |
| SAMD9 | UP | | UP |
| TRIM14 | | | |

[a]CSB-PGBD3 fusion protein
[b]UP indicates a difference of 2-fold or more when fusion, CSB, or CSB + fusion is compared to the UVSS1KO control.

**Table 2-4. CSB and CSB-PGBD3 fusion protein induce closely related antiviral responses in different genetic backgrounds.**

When analyzed using MSigDB instead of L2L, 15 interferon-related genes are upregulated by expression of CSB in the CSB compound heterozygote CS1AN line, 12 of which are also upregulated by expression of CSB-PGBD3 fusion protein in the CSB-null UVSS1KO line (Table 2-S4B, highlighted in blue), and 11 of which belong to 3 related functional themes — viral RNA recognition, protein degradation, and membrane-mediated antiviral activities. (A) Regulation of the 15 genes (*leftmost column*) in the seven interferon-related MSigDB datasets. The correlation is mainly in the UP direction except for the dataset of [82] for STAT3 which often suppresses the interferon

response. (B) Regulation of 12 of the 15 genes in the UVSS1KO datasets (Table 2-S1). CSB-PGBD3 fusion protein is denoted as "fusion" in the tables below. UP and DOWN indicate a >2-fold difference in gene expression when CSB, fusion, or CSB + fusion are compared to the tags-only control). For convenience, vignettes of the 15 genes are provided in Appendix 2.

| Cell Line | | | CSB-PGBD3 + CSB | | CSB-PGBD3 | | CSB | | CSB-PGBD3 + CSB *vs* CSB-PGBD3 | | CSB-PGBD3 + CSB *vs* CSB | | CS1AN + CSB | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Direction | | Down | Up | Down | Up | Down | Up | Down | Up | Down | Up | Down | Up |
| | | Count | 357 | 947 | 305 | 259 | 201 | 109 | 383 | 1218 | 72 | 706 | 204 | 110 |
| CSB-PGBD3 + CSB | Down | 357 | | 7 | 30 | 16 | 88 | 1 | 98 | 6 | 23 | 11 | 15 | 4 |
| | Up | 947 | 0.99 | | 48 | 56 | 2 | 76 | 7 | 544 | 3 | 412 | 26 | 13 |
| CSB-PGBD3 | Down | 305 | **8E-15** | **1E-14** | | 2 | 39 | 3 | 3 | 179 | 1 | 51 | 15 | 1 |
| | Up | 259 | **7E-06** | **2E-23** | 0.87 | | 11 | 20 | 120 | 8 | 9 | 16 | 9 | 20 |
| CSB | Down | 201 | **2E-100** | 0.99 | **2E-32** | **3E-05** | | 1 | 46 | 8 | 0 | 34 | 8 | 8 |
| | Up | 109 | 0.83 | **3E-77** | 0.19 | **3E-18** | 0.63 | | 1 | 51 | 7 | 10 | 4 | 2 |
| CSB-PGBD3 + CSB | Down | 383 | **3E-85** | 0.99 | 0.9 | **2E-136** | **5E-37** | 0.85 | | 0 | 28 | 9 | 16 | 15 |
| *vs* CSB-PGBD3 | Up | 1218 | 0.99 | **0** | 2E-140 | 0.97 | 0.87 | **1E-34** | 1 | | 3 | 423 | 32 | 3 |
| CSB-PGBD3 + CSB | Down | 72 | 2E-22 | 0.59 | 0.63 | 3E-07 | 1 | 9E-08 | 2E-30 | 0.76 | | 3 | 3 | 0 |
| *vs* CSB | Up | 706 | 0.59 | 0 | 8E-22 | 0.01 | 3E-15 | 0.003 | 0.86 | 0 | 0.41 | | 20 | 4 |
| CS1AN + CSB | Down | 204 | **2E-06** | **9E-07** | **2E-07** | **0.001** | 0.003 | 0.02 | 6E-07 | 8E-08 | 0.03 | 9E-06 | | 1 |
| | Up | 110 | 0.1 | **0.001** | 0.78 | **4E-18** | **5E-05** | 0.1 | 9E-10 | 0.94 | 1 | 0.46 | 0.73 | |

CSB and CSB-PGBD3 are both required for expression of many genes

CSB reverses the effect of CSB-PGBD3 on gene expression

CSB Up reverses CSB-PGBD3 Down

CSB controls many genes independently of CSB-PGBD3

CSB controls many genes independently of CSB-PGBD3

CS1AN + CSB partly resembles UVSS1KO + CSB-PGBD3

**Table 2-5. A direct comparison of the CSB-null UVSS1KO microarray datasets with the CSB compound heterozygote CS1AN datasets**. Using gene lists generated by submitting probes to MSigDB ([83] we wrote Perl scripts to calculate the number of overlaps between microarray datasets from UVSS1KO-derived pools stably expressing CSB, CSB-PGBD3 fusion protein, both proteins, or tag only (Table 2-1) and the two compound heterozygote CS1AN-derived lines stably expressing CSB or eGFP ([6]; also see Table 2-1). Only genes that were robustly regulated by 2-fold or more relative to the UVSS1KO tag-only and CS1AN eGFP-expressing controls were compared. We also compared the UVSS1KO-derived CSB, CSB-PGBD3, and CSB + CSB-PGBD3 datasets directly to each other to rule out any potential bias introduced by comparing these three datasets to the same tag-only control (Table 2-1). Small boxes below the diagonal give binominal p-values for overlaps between the corresponding expression array datasets calculated using Excel; small boxes above the diagonal give the number of genes responsible for the overlap. Color outlines denote cognate sets of p-values (below the diagonal) and gene counts (above the diagonal) illustrating some of the conclusions that can be drawn from the microarray comparisons. Unexpectedly, expression of CSB in the CS1AN compound heterozygote mostly closely resembles expression of CSB-PGBD3 fusion protein in the CSB-null UVSS1KO line (purple outlines).

## Supplementary Table Legends

For the sake of brevity, the content of these tables is not included in the text of this document. They can be accessed as Supplementary Materials of the online version of Bailey, et al., 2012 [26]. The labels vary in the online version by omission of the "2-" prefix used here to denote that these tables are part of Chapter 2 of this thesis.

**Table 2-S1.** Gene expression changes resulting from stable expression of CSB, the CSB-PGBD3 fusion protein, or both proteins in CSB-null UVSS1KO cells. (A) A complete list of all robust gene expression changes, defined as increases or decreases of 2-fold or more, resulting from stable expression in CSB-null UVSS1KO cells of CSB, CSB-PGBD3 fusion protein, or both proteins compared to the tags-only control (Table 2-1). For genes with multiple probes, at least one of which exhibited a robust change, the probe with the most extreme fold change is shown in (a) whereas all probes with changes of 2-fold or more are shown in (b). Changes of less than 2-fold are indicated by dashes. Although all expression array data are presented using 2 decimal places, 2-fold cutoffs were made using 3 significant figures. Fold changes are color coded green (up) and red (down), and highlighted according to the cell lines that regulate them: CSB + fusion, CSB, and fusion (orange highlight); CSB + fusion and CSB (pink); CSB only (lavender); CSB + fusion and fusion (yellow); CSB and fusion (light blue); fusion only (dark blue); or CSB + fusion only (light green). (B) The robustly regulated gene lists in S1A(a) were culled to include only those genes that fall into at least one of the 10 functional categories derived by L2L analysis and indicated in columns F through O of panels (a) and (b) based on the gene description in the NCBI Gene Database. (a) Genes are grouped and highlighted according to the cell lines that regulate them as in Table 2-S1A(a). Genes marked by an asterisk are one of several related genes, pseudogenes, or splice variants that react with a single Affymetrix probe. Note that ERCC6 (aka CSB) and PGBD3 are both increased, as expected, in cells expressing cDNAs for CSB + fusion or CSB alone, but not in cells expressing cDNA for the fusion protein only. (b) As in (a) but genes are listed alphabetically. (c) As in (b) but including only genes compiled by searching the NCBI Human Gene Database with the key phrases "interferon regulated," "interferon induced," and "interferon repressed." Column A was generated by an analogous search for each interferon individually (α, β, γ, ε, κ, and ω). (d) As in (c) but using key phrases for STAT1. (e) As in (c) but using key phrases for any STAT. (f) As in (c) but using key phrases for the immune response. (g) As in (c) but using key phrases for inflammation. (h) As in (c) but using key phrases for regulation of apoptosis. (i) As in (c) but using key phrases for induction of apoptosis. (j) As in (c) but using key phrases for neural development. (k) As in (c) but using key phrases for neural growth. (C) interferon-related genes whose >2-fold induction by the CSB-PGBD3 fusion protein is reversed by coexpression of the CSB protein. The fold inductions are highlighted as in S1B(a) and S1B(b). Although IFIT3 (1.70-fold induced by fusion over CSB) and PMAIP1 (1.64-fold induced by fusion over CSB) both fell short of the 2-fold cutoff, IFIT3 was included along with IFIT1, 2, and 5 for completeness, and PMAIP1 was included because it is induced in many other interferon-related datasets (Table 2-5A). Highlighting as in Table 2-S1A(a).

**Table 2-S2.** RT-PCR quality test of expression microarray data. RT-PCR was performed in triplicate as described [6] for both for the gene of interest (e.g. ADA) and for the GAPDH internal control using the same three RNA preparations as in the microarray analysis (Table 2-S1). The $C_t$ values for the triplicate experimental and control reactions were averaged to give $C_t$avg(ADA) and $C_t$avg(GAPDH), and the control was subtracted from the experimental [$C_t$avg(ADA) – $C_t$avg(GAPDH)] to give $\Delta C_t$(ADA) where $C_t$ is the number of PCR cycles required to reach the inflection point for sigmoidal amplification, and $2^{DC_t}$ is the fold change for expression of the gene of interest relative to the control. The $\Delta C_t$(ADA) values for three RNA samples from each cell pool were averaged to give $\Delta C_t$(ADA)avg for that pool. The $\Delta C_t$(ADA)avg for two different pools was subtracted $\Delta C_t$(ADA)avg (CSB) – $\Delta C_t$(ADA)avg (FLAG-HA) to give $\Delta C_t$ for the fold change of the gene of interest in the experimental pool relative to the control pool.

**Table 2-S3.** The CSB-PGBD3 fusion protein induces many of the same genes as an unphosphorylatable U-STAT1 mutant (Y701F-STAT1 aka YF-STAT1). (a) The 52 genes regulated by unphosphorylatable YF-STAT/U-STAT1 in normal BJ fibroblasts, or (b) the 25 genes regulated by overexpression of STAT1 in BJ cells, are compared with genes regulated by 2-fold or more when CSB, the CSB-PGBD3 fusion protein, or both proteins are expressed in UVSS1KO cells (Table S1A). Forced expression of U-STAT1 in normal BJ fibroblasts is known to mimic the late stages of an interferon response in which IFN-related gene expression is driven by unphosphorylated STATs [41]. HyeonJoo Cheon and George Stark (Lerner Research Institute and Department of Genetics, Case Western Reserve University) kindly provided the complete list of U-STAT1 regulated genes from which the core list was derived [41].

**Table 2-S4.** A search for additional functional signatures in the UVSS1KO datasets. (A) We used the curated microarray database MSigDB (www.broadinstitute.org/gsea/msigdb/index.jsp and [83]) to identify the 50 microarray expression datasets in the Chemical and Genetic Perturbations subset of MSigDB that best matched the 259 genes upregulated by 2-fold or more when the CSB-PGBD3 fusion protein is expressed in a CSB-null UVSS1KO background (Table 2-S1). Of these 50 datasets, 47 could be sorted into 4 broad categories (6 chromatin, highlighted in yellow; 19 cancer, orange; 15 interferon response, aqua; and 7 inflammation, green). In addition to a list of the top 50 datasets, MSigDB also returns a p-value for the number of genes in each overlap (Gene Set Names, *upper panel*) as well as a list of all genes in the 50 overlaps (Gene/Gene Set Overlap Matrix, *lower panel*). These genes are listed in groups corresponding to the highest ranking dataset that contains them, and within each group in order of aggregate representation in the highest ranking datasets. (B) The genes in the 50 overlaps were ranked according to the number of datasets containing each gene, and then scored for the number of these datasets that fell into each of the 4 broad functional categories. Genes that are strongly regulated by U-STAT1 in the prolonged interferon response (Table 2-3) are shown in bold red; and interferon-related genes that are upregulated both by expression of CSB in CS1AN and by expression of

CSB-PGBD3 fusion protein in UVSS1KO (Table 2-5) are highlighted in blue. The rank order of the genes in each of the 4 categories is similar.

# Chapter 3:
# Tethering of the conserved piggyBac transposase fusion protein CSB-PGBD3 to chromosomal AP-1 proteins regulates expression of nearby genes in humans

## Summary

The CSB-PGBD3 fusion protein arose over 43 million years ago when a 2.5 kb piggyBac 3 (PGBD3) transposon inserted into intron 5 of the Cockayne syndrome Group B (CSB) gene in the common ancestor of all higher primates. As a result, full length CSB is now coexpressed with an abundant CSB-PGBD3 fusion protein by alternative splicing of CSB exons 1-5 to the PGBD3 transposase. An internal deletion of the piggyBac transposase ORF also gave rise to 889 dispersed, 140 bp MER85 elements which were mobilized in trans by PGBD3 transposase. The CSB-PGBD3 fusion protein binds MER85s *in vitro*, and induces a strong interferon-like innate antiviral immune response when expressed in CSB-null UVSS1KO cells. To explore the connection between DNA binding and gene expression changes induced by CSB-PGBD3, we investigated the genome-wide DNA binding profile of the fusion protein. CSB-PGBD3 binds to 363 MER85 elements *in vivo*, but these sites do not correlate with gene expression changes induced by the fusion protein. Instead, CSB-PGBD3 is enriched at AP-1, TEAD1, and CTCF motifs, presumably through protein-protein interactions with the cognate transcription factors; moreover, recruitment of CSB-PGBD3 to AP-1 and TEAD1 motifs correlates with nearby genes upregulated by CSB-PGBD3 expression in UVSS1KO cells and downregulated by CSB rescue of mutant CS1AN cells. Consistent with these data, the N-terminal CSB domain of the CSB-PGBD3 fusion protein interacts with the AP-1 transcription factor c-Jun and with RNA polymerase II, and a chimeric CSB-LacI construct containing only the N-terminus of CSB upregulates many of the genes induced by CSB-PGBD3. We conclude that the CSB-PGBD3 fusion protein substantially reshapes the transcriptome in CS patient CS1AN, and that continued expression of the CSB-PGBD3 fusion protein in the absence of functional CSB may affect the clinical presentation of CS patients by directly altering the transcriptional program.

# Introduction

Cockayne syndrome (CS) is a neurodevelopmental disorder most often caused by loss of functional CSB or CSA protein (OMIM #133540 or #216400) [2]. CSB is a SWI/SNF2-like ATPase and chromatin remodeling protein that plays a key role in transcription-coupled nucleotide excision repair (TC-NER) of helix-distorting DNA lesions. When RNA polymerase II (RNAPII) stalls at a site of DNA damage, CSB is among the first proteins to bind [4,84,85] and is required to recruit other NER factors including CSA and the TFIIH complex containing the XPB and XPD helicases [8,9,86]. CSB is also known to activate RNA polymerase I (RNAPI) transcription of ribosomal RNA [3], and to induce changes in gene expression resembling those caused by chromatin remodeling and histone modification [6].

We recently discovered a domesticated PGBD3 transposon (piggyBac transposable element-derived 3) that inserted into intron 5 of the CSB gene at least 43 Mya in the common ancestor of marmoset and humans. As a result, primate CSB genes including our own now generate both full length CSB (coding exons 2-21) and — by alternative splicing and polyadenylation — a CSB-PGBD3 fusion protein that joins the N-terminal domain of CSB (coding exons 2-5) to the intact PGBD3 transposase [14]. CSB-PGBD3 is startlingly well conserved from marmoset to humans, whereas four other identifiable copies of the PGBD3 transposon elsewhere in the human genome have all decayed into pseudogenes (PGBD3P1-4). The PGBD3 transposon contains a 5’ splice acceptor site just upstream of the transposase ORF and a polyadenylation signal downstream of the ORF that allow alternative splicing of CSB exon 5 to the intact transposase without precluding continued expression of full length CSB (Figure 3-1). In fact, the insertion of PGBD3 expanded the repertoire of the CSB locus from one protein to three: full length CSB, the more abundant CSB-PGBD3 fusion protein, and most abundant of all, the intact PGBD3 transposase transcribed from a cryptic promoter near the 3’ end of CSB exon 5 [14]. Coexpression of the CSB-PGBD3 fusion protein with CSB initially suggested that the fusion protein might contribute to or modulate CS disease [14]; however, mutations that cause CS are distributed across the entire length of the CSB gene (except in the PGBD3 transposon) and no consistent clinical differences have been observed between CS patients with CSB mutations in coding exons 2-5 (many of whom do not make the CSB-PGBD3 fusion protein) and patients with mutations in exons 6-21 (who continue to make the CSB-PGBD3 fusion protein) [2].

Unlike the ATPase domain (CSB exons 6-21), the function of the N-terminal domain (coding exons 2-5) shared by CSB and the CSB-PGBD3 fusion protein is not yet well understood (Figure 3-1). The only recognizable motif in exons 2-5 is a highly acidic domain between E356 and E403 containing 25 aspartates and glutamates, but this

domain does not appear to be essential for recovery of RNA synthesis following UV damage [87,88]. Interestingly, the N-terminus autoinhibits association of CSB with chromatin in both normal and UV-irradiated cells, and ATP hydrolysis is required for relief of inhibition [89]. The isolated N-terminal domain has also been shown to interfere with transcription and repair: Truncated CSB protein expressed in the patient-derived cell line CS1AN represses elongation by RNAPI [90] and the N-terminus of CSB interacts with topoisomerase I (Top1) to inhibit repair of Top1 adducts both as part of the CSB-PGBD3 fusion protein and independently [37].

We have recently shown that expression of the CSB-PGBD3 fusion protein in CSB-null UVSS1KO cells induces a strong transcriptional response dominated by an interferon-like innate antiviral immune response that may be driven by upregulation of the STAT1, STAT2, and IRF9 components of the heterotrimeric transcription factor ISGF3 (interferon-stimulated gene factor 3) [26]. As might be expected from conservation of the CSB-PGBD3 fusion protein for over 43 My, the interferon-like response induced by CSB-PGBD3 is dramatically repressed by coexpression of full-length CSB, and is not induced by CSB alone. However, the mechanism by which the CSB-PGBD3 fusion protein induces the interferon-like response, and CSB represses it, are still unclear.

The CSB-PGBD3 fusion protein may affect RNAPII gene expression through both global and local mechanisms. Globally, CSB-PGBD3 may modulate CSB functions by interacting with complexes that normally contain functional CSB; this could explain how the fusion protein modulates DNA repair without inducing or repressing transcription of known DNA repair factors [26]. CSB-PGBD3 may also affect RNAPII transcription locally by binding to dispersed DNA elements called MER85s, thereby regulating expression of nearby genes.

PGBD3, like many autonomous mobile elements, has given rise to a family of internally-deleted, nonautonomous elements that can be mobilized by the PGBD3 transposase. These 140 bp MER85 elements retain about 100 bp from the 5’ end of PGBD3, and about 40 bp from the 3’ end, but have lost the transposase ORF along with the upstream 5' SS and the downstream poly(A) site. We have identified 889 MER85 elements dispersed throughout the human genome, most of which include 13 bp terminal inverted repeats (TIRs) that are required by the PGBD3 transposase for excision and reinsertion into TTAA target sites. We have also demonstrated that MER85 elements bind PGBD3 and CSB-PGBD3 *in vitro* [26]. Thus, CSB-PGBD3 may enable MER85s to recruit the N-terminus of CSB to specific genomic loci where it can affect local chromatin structure or recruit transcription and repair factors.

We wish to understand why the CSB-PGBD3 fusion protein is so well conserved, and to determine what roles it may play in health and CS disease. Here, we explore the connection between the genome-wide DNA binding profile of CSB-PGBD3 and transcriptional regulation in UVSS1KO cells. As expected, we find that CSB-PGBD3 binds directly *in vivo* to many MER85 elements throughout the genome. Surprisingly, we also find that CSB-PGBD3 binds indirectly to TRE motifs (tumor promoting antigen response elements) recognized by AP-1 family (activating protein-1) transcription factors, as well as to motifs for the TEAD1 (TEA domain family member 1) and CTCF (CCCTC-binding factor) transcription factors. We show that CSB-PGBD3 physically interacts with the AP-1 protein c-Jun, and that genes upregulated by CSB-PGBD3 correlate with binding of CSB-PGBD3 to nearby TRE motifs but not with binding to MER85 elements. We also show that CSB-PGBD3 interacts with RNAPII (RNA polymerase II), and that interactions with RNAPII and c-Jun are both mediated primarily by the N-terminal CSB domain of CSB-PGBD3. Thus despite the ability of the CSB-PGBD3 fusion protein to bind specifically to MER85s both *in vitro* and *in vivo*, binding does not appear to have widespread transcriptional consequences. In contrast, binding of the CSB-PGBD3 fusion protein to TRE motifs through protein-protein interactions with c-Jun and possibly other AP-1 family members correlates with genes involved in angiogenesis [91,92], innate immunity [93], and the Smad2/3 and TGF-beta pathways [94], demonstrating that the CSB-PGBD3 protein modulates a preexisting AP-1-based regulatory network. Whether these regulatory effects were responsible for initial fixation of the CSB-PGBD3 fusion protein in the common ancestor of humans and marmoset 43 Mya, or whether these regulatory effects have evolved over time, remains to be seen.

# Results

### The CSB-PGBD3 fusion protein binds to the 5’ end of MER85 elements *in vivo*

Using a panel of six highly conserved MER85s with >90% identity to the Repbase MER85 consensus [95], we found previously that both the CSB-PGBD3 fusion protein and solitary PGBD3 transposase can bind MER85 elements *in vitro* [26]. To extend these results to living cells, we performed ChIP-PCR (chromatin immunoprecipitation followed by radiolabeled PCR) using human euploid HT1080 fibrosarcoma cells, genomic primers for the same six MER85 elements, and antibodies directed against the N- or C-terminus of CSB (Figure 3-2). We first confirmed that in HT1080 cells, which are wild-type for CSB and CSB-PGBD3, antibody against the N-terminus of CSB immunoprecipitated both CSB and CSB-PGBD3, whereas antibody against the C-terminus brought down CSB alone (data not shown). ChIPs with antibody against the N-terminus of CSB enriched for 5 of 6 MER85 elements *in vivo* including all 4 elements that shifted in the electrophoretic mobility shift assay (EMSA) [26]; ChIPs using antibody

against the C-terminus or nonspecific antibody did not enrich for any of the six MER85s (Figure 3-2).

To explore the DNA sequence requirements for CSB-PGBD3 binding to MER85 elements, we performed EMSAs with two strongly bound MER85 elements (MER85-360 and MER85-427, see Table S1) that contain the 13 bp TIR sequences required for transposition. Surprisingly, when the MER85s were cut in two at the unique DpnI site, only fragments containing the 5'-most 42 bp of MER85 sequence exhibited a mobility shift (Figure 3-3). We confirmed this result by EMSAs using synthetic 42-mers that corrected occasional mismatches between the two MER85s and the consensus MER85 sequence (Figure 3-4). Thus the TIR sequence is not sufficient for binding the PGBD3 transposase, and essential sequences of the transposase binding site must be located elsewhere within the 5'-most 42 bp of MER85 elements.

**An imperfect internal 16 bp palindrome is essential for binding of the CSB-PGBD3 fusion protein to the 5' end of MER85 elements *in vitro***

Visual inspection of MER85 sequences revealed an imperfect 16 bp palindrome GTTCCAtTAtTGGAAC located 3 bp internal to the 5' TIR. The PGBD3 transposon that integrated into the CSB gene contains the same palindrome at three locations: once near the 5' TIR as in MER85s, again 59 bp upstream of the PGBD3 transposase ORF, and yet again 75 bp downstream of the ORF termination codon and 114 bp upstream from the 3' TIR (Figure 3-5; also see Figure 3-6 for conservation of the palindromes in PGBD3 pseudogenes). In MER85 elements, the sole palindrome lies 3 bp downstream from the 5' TIR but 96 bp upstream of the 3' TIR. Similar spacing between the 3' most palindrome and the 3' TIR in both the PGBD3 transposon (114 bp) and MER85s (96 bp) suggests that the sole MER85 palindrome may be functionally equivalent to the 3' most palindrome in the full-length transposon, or may perhaps do double duty — functioning early in the reaction at the 5' end and later at the 3' end. A similar palindrome TGCGTaAAATTgACGCA, called the internal repeat, is found 3 bp downstream from the 5' TIR and 31 bp upstream from the 3' TIR of the piggyBac transposon from *Trichoplusia ni* [96]. A partial deletion of the 3' internal repeat abolishes transposition [97], suggesting that the palindromes are functionally important for transposition by both the moth and human piggyBac elements.

To determine whether the 5' palindrome of MER85s is required for PGBD3 binding, we examined MER85-65 in greater detail. This was the only MER85 in the panel of 6 that did not bind PGBD3 transposase *in vitro* or *in vivo*, despite being nearly identical in sequence to the other 5 elements [26]. Inspection of the 5' end of MER85-65 revealed mismatches at 4 positions compared to the MER85 consensus: 2 in the TIR, and 2 in the palindrome (Figure 3-7). To test if the mutations in the 5' TIR or the palindrome or

both reduced the binding affinity, we performed EMSAs with 42 bp oligonucleotides that contained these mutations, singly and in combination, but otherwise matched the 5’ end of the MER85 consensus. Oligonucleotides with mutations that matched the MER85-65 TIR exhibited little loss of binding compared to the consensus. In contrast, oligonucleotides with mutations that matched the MER85-65 palindrome exhibited a 60% loss of binding (Figure 3-7), suggesting that mutations in the palindrome are likely responsible for the lack of binding to this particular element *in vivo*. No combination of mutations in the oligonucleotide gave as great a loss of binding as observed when the entire MER85-65 element was assayed by EMSA [26] or ChIP-seq (Figure 3-2), suggesting that other factors, such as sequence context or chromatin accessibility, may contribute to CSB-PGBD3 binding *in vivo* [R1]. To confirm the importance of the 5' palindrome, we tested 42 bp oligonucleotides in which either the entire 5' TIR or 5' palindrome was replaced by random sequence. Surprisingly, deletion of either region reduced binding *in vitro*, but the effect was greater for the palindrome (80% loss) than for the TIR (60% loss) (Figure 3-7).

The fact that 5' MER85 sequences favor DNA binding both *in vitro* (Figure 3-3) and *in vivo* (Figure 3-2, Table S1) suggests that the PGBD3 transposase alone is sufficient for initial recognition of the 5’ end of MER85 mobile elements. The ability of the moth element to function efficiently in mammalian cells further reinforces this interpretation [98]; however, host independence does not exclude the participation of auxiliary proteins that may facilitate or stabilize assembly of the transpososome [99].

**The CSB-PGBD3 fusion protein is enriched at >2,000 sites in the human genome**

CSB-null UVSS1KO fibroblasts are derived from a patient with UV sensitive syndrome (UVSS) and express neither CSB [15] nor CSB-PGBD3 fusion protein [14] as a result of a homozygous nonsense mutation at CSB codon 77. We had previously generated gene expression array data for UVSS1KO cells stably expressing FLAG-HA-tagged CSB-PGBD3 fusion protein [26]. To correlate these expression array data with genome-wide CSB-PGBD3 chromatin binding profiles for the same cells, we used paired-end ChIP-seq [100] in which the cells are crosslinked with formaldehyde, sonicated, and sheared chromatin is immunoprecipitated with an antibody against the protein of interest — in this case a mouse monoclonal antibody against the N-terminal domain of human CSB. The immunoprecipitated DNA fragments are ligated to Illumina adapters, and 300-600 bp fragments are size-selected by PAGE and pre-amplified by PCR before loading onto the Illumina flow cell where one end of each captured fragment is sequenced. Synthesis of the opposite strand and cleavage of an 8-oxoguanine incorporated into the immobilized flow cell oligonucleotides then allow the fragments on the surface of the flow cell to be resequenced from the other end [100]. Paired-end sequencing greatly improves the mapping of repetitive DNA sequence elements such as MER85s because

the short reads obtained from both ends of each sonicated chromatin fragment can be required to align uniquely with genomic sequences near each other and on opposite strands.

More than 8.5 million pairs of enriched ChIP-seq reads of 36 bp were mapped to human genome build hg18 (NCBI 36) using the read mapping program Bowtie [101]. Because CSB-PGBD3 binds to repetitive (and very similar) MER85 elements, we used stringent settings that disregard reads containing mismatches and reads that could not be uniquely mapped. The surviving reads were then analyzed for local enrichment using three independent peak-finding algorithms — MACS [102], ERANGE [103], and QuEST [104] — which differ based on how the paired sequence tags are handled, as well as in the statistical methods used to determine peak enrichment (reviewed in [105]). Comparison of results from each algorithm allowed us to find peaks that were consistently enriched independent of the peak-calling method.

We found that 363 of 889 MER85 elements were reliably enriched and called as peaks by all 3 peak finding algorithms (Table S1). To prevent easily sheared chromatin regions and regions artefactually enriched by pre-amplification from scoring as peaks, each of our analyses included an input control consisting of ~3 million single-end reads from the same sheared chromatin used for ChIP-PET. The 2,087 peaks found by all 3 algorithms were used for subsequent analysis (Table S2). We then wrote a Perl script to generate internally consistent CSB-PGBD3 binding profiles over all 2,087 peaks. The script converted mapped paired-end reads to the genomic coordinates of the corresponding ChIP fragments, calculated the number of fragments overlapping each position in the genome, and compiled the fragment map as a WIG file to display and analyze CSB-PGBD3 binding profiles (Appendix 3). A second script was used to locate the highest fragment overlap, defined as the peak summit, in each of the 2,087 enriched region identified by all three peak calling algorithms. We also used the Cis-regulatory Element Annotation System (CEAS) program [106] to show that CSB-PGBD3 peaks are significantly enriched within 3 kb of transcription start sites (6.1%, p-value 1.6e-20), although the vast majority of peaks are either intronic (41.6%) or in distal intergenic regions (47.8%) (Figure 3-8).

**The CSB-PGBD3 fusion protein binds to the 5’ end of 363 MER85 elements and the PGBD3 locus in CSB**

We located all 889 MER85 elements in the hg18 build of the human genome, and examined them individually to ensure that the boundaries of each MER85 were correctly identified even in cases where expansions or insertions altered the length of MER85 elements. All MER85 elements are given in the same orientation as the parental PGBD3 transposon (Figure 3-1) with the 5’ and 3' ends of the MER85s corresponding to

the first ~100 bp and last ~40 bp of the transposon. Of these 889 MER85s, we found 813 with intact terminal inverted repeats (TIRs), 13 of which had large internal insertions or repeat expansions (>20 bp longer than normal) and 1 of which had an internal deletion. Of the remaining 76 MER85s, 22 had incomplete 5’ ends, 49 incomplete 3’ ends, and 5 lacked both TIRs (Table S1).

When all bound MER85s were aligned in the same orientation, fragment overlaps indicated preferential binding to a 40 bp region just internal to the 5' TIR (Figure 3-9), consistent with EMSA experiments on representative MER85 elements (Figure 3-3). No MER85 element lacking the 5’ palindrome bound CSB-PGBD3, although several elements that lacked 5’ or 3’ TIRs were reliably enriched (Table S1), further supporting our conclusion that CSB-PGBD3 binds primarily to the 5' palindrome. Comparison of the 5’ palindrome sequences of bound and unbound MER85s revealed that 291 of 363 bound elements (80.1%) but only 48 of 526 unbound elements (9.1%) perfectly matched the consensus. The presence of unbound MER85s with perfect palindrome sequences suggests once again that other factors, such as chromatin accessibility, are likely to modulate CSB-PGBD3 binding *in vivo*. This could also explain why only 5 of the 6 MER85 elements used previously for EMSA correlated with the ChIP-PCR and ChIP-seq results (Figure 3-2).

We also examined binding of the CSB-PGBD3 fusion protein to the PGBD3 locus within the CSB gene, as well as to PGBD3 pseudogenes. Unexpectedly, the PGBD3 locus in CSB is one of the strongest and most extensive CSB-PGBD3 binding sites in the entire genome; moreover, paired-end fragment reads overlapped most heavily near each of three copies of the imperfect 16 bp palindromic sequence in the PGBD3 transposon (Figure 3-5). The same was true for the PGBD3 pseudogenes (Figure 3-6), but only where the palindromic repeats perfectly matched those of the full-length PGBD3 insertion in CSB (Figure 3-6). Although CSB is thought to be expressed in all tissues, and CSB mutations are recessive, it is unclear if or how binding of the CSB-PGBD3 fusion protein to the PGBD3 transposon affects CSB and/or PGBD3 transcription, splicing, or expression.

**TRE, TEAD1, and CTCF motifs are enriched in CSB-PGBD3 peaks**

Much to our surprise, peaks over MER85, PGBD3, and PGBD3 pseudogenes accounted for only 367 (17.5%) of the 2,087 genomic regions enriched by immunoprecipitation with CSB-PGBD3. To determine what sequences in non-MER85 peaks were responsible for enrichment of CSB-PGBD3, we used MEME (Multiple Em for Motif Elicitation) to search for overrepresented sequence motifs located within 50 bp of non-MER85 peak summits [107]. Enriched motifs were then submitted to the TOMTOM motif comparison tool to identify known binding proteins [108].

The top hit was the sequence TGANTCA found near 585 (28%) of the 2,087 peak summits (E-value = 6.4e-335) (Figure 3-10). This motif was identified by TOMTOM as the tumor promoting antigen response element (TRE) best known as the binding site for Activator Protein 1 (AP-1) family complexes [109].

The next most highly represented motif was [AT]GGAAT[GT] where [AT] is A or T, and [GT] is G or T; this motif is found near 269 (13%) of the 2,087 peak summits (E-value = 3.1e-64) and resembles the binding site for the TEAD1 (TEA domain family member 1) transcription enhancer protein (Figure 3-10). This motif is very similar to part of the MER85 palindromic region (TGGAACG), and we cannot entirely exclude the possibility that it is bound directly by CSB-PGBD3 because a C>T mutation within this motif (TGGAATG) only slightly reduced PGBD3 binding *in vitro* (Figure 3-7). On the other hand, 199 MER85 elements in the genome have this C>T mutation, yet only 6 are bound by CSB-PGBD3 in the ChIP-seq dataset (Table S1).

The third most significant motif with a known binding protein was CCA[CG][CT]AG[AG][GT]GGC, found near 58 (2.7%) of the 2,087 peak summits (E-value 1.6e-9) and was identified as the binding site for CTCF (CCCTC-binding factor), a key regulator of chromatin looping and other higher-order chromatin structures [110] (Figure 3-10).

The overrepresentation of these three motifs near the CSB-PGBD3 summits in non-MER85 peaks (tabulated in Table S3) suggests that CSB-PGBD3 may interact with all three of these DNA binding factors. Consistent with this interpretation, average fragment overlap profiles centered on these motifs show sharp accumulation of CSB-PGBD3 enriched fragments over the motifs (Figure 3-11). Alternatively, CSB-PGBD3 might bind directly to one or more of these motifs, for example through a cryptic activity of PGBD3 DNA binding domain.

**The AP-1 family protein c-Jun co-immunoprecipitates with the CSB-PGBD3 fusion protein**

We used an EMSA assay to ask whether CSB-PGBD3 can bind directly to TRE motifs, or is more likely tethered to the motif by protein-protein interactions with TRE bindng factors. As anticipated, purified CSB-PGBD3 fusion protein failed to shift 42 bp oligonucleotides containing one or two TRE motifs, although control MER85 sequences shifted cleanly and random sequences did not shift at all (Figure 3-4).

To determine if binding of CSB-PGBD3 to TRE motifs is mediated by an interaction with a TRE binding protein, we asked whether CSB-PGBD3 would co-immunoprecipitate

(coIP) with AP-1 proteins that are known to bind TRE motifs. AP-1 complexes are composed of many homo- or heterodimeric combinations of members of the Jun, Fos, Maf, and ATF protein families, and the combination of AP-1 family members determines the affinity of the complex for specific variants of the sequence motifs [93,111]. Fos and Jun bind preferentially to the TRE sites (TGANTCA) identified in CSB-PGBD3 peaks, and more weakly to the similar cyclic AMP response element binding site (TGACGTCA). Although the binding repertoire of Jun and Fos can be expanded through interactions with several other DNA binding proteins [112], the CSB-PGBD3 peaks contain only TRE motifs suggesting that CSB-PGBD3 interacts directly with Jun or Fos proteins.

The Jun and Fos genes c-Jun, JunD, Fra1, and Fra2 have previously been shown to be expressed in exponentially growing fibroblast cultures [113]. We were able to detect expression of Jun, JunD, and Fra2 in our UVSS1KO-derived fibroblast lines by Western blotting (Figure 3-13) but not Fra1 (data not shown). In UVSS1KO cells stably expressing FLAG-HA-tagged CSB-PGBD3, coIPs with antibodies against c-Jun enriched for CSB-PGBD3 compared to a non-specific antibody control (Figure 3-12, right panel) but coIPs with antibodies against JunD and Fra2 did not (data not shown). Moreover, reciprocal coIPs with anti-FLAG antibodies enriched for c-Jun in cells expressing FLAG-HA-CSB-PGBD3 (Figure 3-12A). These results suggest that CSB-PGBD3 binds to TRE sites indirectly, through a protein-protein interaction with bound c-Jun.

To localize the site of interaction on CSB-PGBD3, we repeated the coIPs in cells expressing FLAG-HA-tagged chimeric CSB-eGFP, eGFP-PGBD3, or full-length CSB (Figure 3-12B). Of these cell lines, only CSB-eGFP enriched for c-Jun in an anti-FLAG coIP. Thus c-Jun interacts with the N-terminus of CSB in the CSB-PGBD3 fusion protein, but not with the N-terminus of intact CSB protein. CSB may fail to bind c-Jun because the autoinhibitory N-terminal domain preferentially interacts with the C-terminal helicase domain in the intact protein [89].

**CSB-PGBD3 binding to TRE motifs, but not MER85s, correlates with regulation of nearby genes in CSB-null UVSS1KO cells, and with CSB repression in CS1AN cells that continue to express the CSB-PGBD3 protein**

We used the Genomic Regions Enrichment of Annotations Tool (GREAT) [40] to ask whether genes that are regulated by the CSB-PGBD3 fusion protein [26] are located near CSB-PGBD3 binding sites as determined by ChIP-seq. We previously generated expression array datasets for stable expression of CSB-PGBD3, CSB, both proteins, or neither in CSB-null UVSS1KO cells [26] but these data had not yet been entered into a database used by the online version of GREAT. Instead, we used a local copy of the

GREAT tool, Calculate Binomial P-Value, to correlate our CSB-PGBD3 expression array and ChIP-seq data. We also compared our CSB-PGBD3 ChIP-seq data to genes up- and downregulated when the CS1AN cell line, a patient-derived CSB compound heterozygote, was rescued with wild-type CSB [6].

GREAT tests for statistical enrichment of peaks in regions near a set of genes. To do this, GREAT defines “regulatory domains” that extend in both directions for a specified distance from the transcription start site (TSS) or to the next nearest gene. Using regulatory domains of 100 kb, 250 kb, and 1 Mb, we tested sets of genes that were up- and downregulated under each condition separately, and compared them to the set of all 2,087 CSB-PGBD3 peaks. We also used GREAT to correlate our expression array datasets with CSB-PGBD3 peaks over MER85 elements (363 peaks), TRE motifs (585 peaks), TEAD1 motifs (269 peaks), CTCF motifs (58 peaks), and peaks that contain none of these motifs (892). Very few peaks contained more than one motif except for 72 peaks with both TRE and TEAD1 motifs, and for consistency these TRE+TEAD1 peaks were counted as members of both peak sets. For each comparison, 100 sets of randomized peak locations were used as negative controls and to calculate empirical false discovery rates (FDR) [114]. Only comparisons with an FDR of less than 1% were considered significant (Table S4).

GREAT analysis revealed that peaks containing TRE motifs are significantly enriched near genes upregulated and downregulated by CSB-PGBD3 using all of the regulatory domain sizes (orange cells in Table S4). Enrichment of TRE motifs near upregulated and downregulated genes suggests that CSB-PGBD3 interacts with AP-1 proteins to modulate the expression of nearby genes. In contrast, peaks over MER85 elements did not correlate significantly with any of the UVSS1KO or CS1AN expression array datasets (gray cells in Table S4), despite enrichment of MER85 elements near specific GO categories [14]. This suggests that regulation of gene expression by the CSB-PGBD3 fusion protein is strongly dependent on location and cooperation with other transcription factors; simple DNA binding in the vicinity of genes is not sufficient. These results support a very different model from our initial speculation that CSB-PGBD3 binding would create a MER85-based transcriptional network. Instead, it appears that CSB-PGBD3 selectively interacts with existing transcription factors to provide an additional layer of gene regulation on top of established regulatory networks.

**CSB-PGBD3 and CSB may coregulate expression of specific genes in normal individuals**

[R1] In addition to analysis of genes regulated by CSB-PGBD3 expression alone, we also compared CSB-PGBD3 binding to genes regulated by coexpression of CSB and CSB-PGBD3 in the same CSB-null cell line UVSS1KO. Importantly, this set of genes is

distinct from genes regulated by CSB or CSB-PGBD3 alone, suggesting that co-regulation could be the result of direct interactions between the N-terminus of CSB-PGBD3 and CSB [89] or indirect interactions in which upregulation of certain genes by CSB-PGBD3 requires prior (or concurrent) chromatin remodeling by CSB [6]. GREAT analysis revealed that many genes which are upregulated by coexpression of CSB and CSB-PGBD3, but not by either protein alone, correlate significantly with the set of all peaks bound by CSB-PGBD3 and with the subsets of peaks over TRE and TEAD1 motifs (blue cells in Table S4). These genes could in principle be regulated by the N-terminal domain of CSB, CSB-PGBD3, or both; however, we might then have expected to see a similar correlation with genes upregulated by stable expression of CSB alone. It therefore seems more likely that this subset of genes is upregulated by CSB-PGBD3 through interactions that are enhanced by or require CSB, and thus may also be upregulated in normal, healthy individuals.

In contrast, many genes that are upregulated by expression of CSB-PGBD3 in CSB-null cells are repressed by coexpression of CSB [26]. Moreover, 16 of these genes are also downregulated (binomial p-value 6e-7) when CSB is expressed in CS1AN cells that continue to express the CSB-PGBD3 fusion protein despite loss of functional CSB [14]. Of these 16 genes, 8 have CSB-PGBD3 binding sites within 100 kb of the TSS (ARHGAP29, IGFBP7, MGLL, PODXL, PSG1, RGMB, RGS4, and SERPINE1) suggesting that CSB can repress some, but not all genes that are upregulated by nearby CSB-PGBD3 fusion protein — perhaps depending on local context or the specific transcription factor(s) that tether CSB-PGBD3 to the site.

**CSB-PGBD3 peaks correlate with diverse ontologies related to angiogenesis, the TGF-beta pathway, cancer, and immune responses**

Our expression array analysis was limited to several cell lines and culture conditions. To investigate the role of CSB-PGBD3 binding sites in the broader context of human biology and disease, we used the online version of GREAT to compare our binding sites to a diverse set of gene ontologies. Using the default settings, we submitted either the full set of CSB-PGBD3 peaks to GREAT, or the subsets containing the MER85, TRE, TEAD1, or CTCF motifs, or no recognizable motif. The MER85 and CTCF (as well as TRE+TEAD1) peaks did not exhibit statistically significant overlaps with any ontology sets, but for the other peak categories we examined the top five results in the GO Biological Processes, Disease Ontology, Pathway Commons, and MSigDB Perturbation datasets (Table S5). We found that CSB-PGBD3 binding sites correlated significantly with genes related to the TGF-beta pathway, carcinogenesis, and IFN and IL-2 driven innate immune responses (see Table S5 legend for details).

**CSB-PGBD3-bound TRE motifs are enriched near CSB-PGBD3-bound MER85 elements**

MER85 elements are among the strongest CSB-PGBD3 binding sites *in vivo*, yet bound MER85 elements do not correlate with genes induced or repressed by CSB-PGBD3 expression in CSB-null UVSS1KO cells (Table S4). Thus we must consider the possibility that continued binding of the CSB-PGBD3 fusion protein to MER85s might be fortuitous or functionless. The burst of MER85 replication apparently came to an end about 35 Mya [115], perhaps upon mutation of the conserved catalytic aspartate (D352) in the PGBD3 transposase ORF to asparagine [14,116]. The limited sequence diversity of the surviving 889 human MER85s (Table S1), the ability of the CSB-PGBD3 binding site to tolerate point mutations and even deletions (Figure 3-7), and the small target size of the essential 16 bp imperfect palindrome (Figures 5 and 6), are all consistent with our observation that at least 40% (363/889) of all MER85s retain the ability to bind the PGBD3 transposase (Table S1) despite ongoing mutations over the past 35 My. We conclude that neutral sequence evolution could have been sufficient to account for the homogeneity and current functions of MER85 elements.

Alternatively, binding of the CSB-PGBD3 fusion protein to MER85s through the PGBD3 domain may enable CSB-PGBD3-mediated chromosome looping with transcription factors bound to TRE, TEAD1, or CTCF motifs. To test this hypothesis, we used the GREAT tools to determine if CSB-PGBD3 binding sites containing TRE, TEAD1, or CTCF motifs (Table S3) were significantly enriched within 100 kb of MER85s that are bound by CSB-PGBD3 (Table S1). Surprisingly, we found a strong correlation between CSB-PGBD3 peaks containing TRE motifs and the 363 MER85 elements bound by CSB-PGBD3 (36 of 585 bound TRE motifs, P-value = 7.9e-7) but not with the 529 unbound MER85 elements (16 of the 585 bound TRE motifs, P-value = 0.88). Peaks containing TEAD1 or CTCF motifs, and peaks containing no identified motif, showed no enrichment near bound or unbound MER85 elements.

**CSB-PGBD3 interacts with RNAPII**

CSB interacts with stalled RNAPII after induction of DNA damage [8], but it also copurifies with RNA polymerase II in unirradiated cells [85] and thus may associate with transcribing RNA polymerase II (RNAPII) as well as with TCR complexes. To determine whether some of the genomic CSB-PGBD3 peaks might reflect interaction of CSB-PGBD3 with RNAPII, we asked if antibody against the C-terminal domain (CTD) of RNAPII could co-immunoprecipitate (coIP) CSB-PGBD3, and vice versa. Intriguingly, coIPs with antibody against the RNAPII C-terminal domain enriched for CSB-PGBD3 but not CSB in undamaged HT1080 human fibrosarcoma cells (Figure 3-14a). In a reciprocal coIP using UVSS1KO cells that stably express FLAG-HA-CSB, FLAG-HA-

CSB-PGBD3, or the FLAG-HA tags only, coIPs with anti-FLAG antibody enriched for RNAPII in cells expressing CSB-PGBD3, but not in cells expressing intact CSB or tags only (Figure 3-14b). The failure of CSB to co-immunoprecipitate RNAPII is more likely to reflect low affinity between CSB and RNAPII in the absence of DNA damage than accessibility of the tags, because anti-FLAG IPs readily pull down FLAG-HA-CSB in UVSS1KO cells (data not shown).

To see if interactions between CSB-PGBD3 and RNAPII could account for CSB-PGBD3 peaks that did not contain an overrepresented sequence motif, we compared regions within 50 bp of CSB-PGBD3 peaks to enriched RNAPII peaks obtained from the Yale TFBS collection in the UCSC Genome Browser database [117,118]. Because RNAPII binding sites vary between cell types, we analyzed 18 RNAPII genome-wide peak sets from 15 cell lines. We found 105 of 2087 CSB-PGBD3 peaks consistently overlapped at least 10 of 18 RNAPII peak sets (Table S6), and that 85 of these CSB-PGBD3 peaks did not contain a MER85, TRE, TEAD1, or CTCF motif. The set of 105 CSB-PGBD3 peaks that overlapped RNAPII peaks were compared to expression array datasets using GREAT as described previously. Peaks associated with RNAPII binding sites were enriched near genes upregulated by coexpression of CSB and CSB-PGBD3, but not by expression of either protein alone (Table S4). Thus, interactions between CSB-PGBD3 and RNAPII may require regulation or remodeling of the gene by CSB [6].

We localized the region of interaction between CSB-PGBD3 and RNAPII by asking whether antibody against the CTD of RNAPII would immunoprecipitate stably expressed FLAG-HA-CSB-eGFP, FLAG-HA-eGFP-PGBD3, or FLAG-HA-CSB-PGBD3 from UVSS1KO cells. Surprisingly, RNAPII interacts with CSB-eGFP but not with eGFP-PGBD3 (Figure 3-14c). Thus, CSB-PGBD3 interacts with RNAPII through the N-terminal CSB domain, just as it does with c-Jun (Figure 3-12). The implication may be that the highly conserved SWI/SNF ATPase domain encoded by CSB exons 6-21, although unlikely to be a generic chromatin remodeler [119], is modulated, autoinhibited [89], and targeted to specific chromosomal locations by the N-terminal domain (coding exons 2-5) (Figures 10 and S1).

**The N-terminal CSB and C-terminal PGBD3 domains of CSB-PGBD3 can independently alter gene expression**

The ability of the N-terminal domain of CSB to interact directly with c-Jun (Figure 3-12) and RNAPII (Figure 3-14), as well as the failure of CSB-PGBD3 fusion protein to affect expression of nearby genes when bound to MER85s (Table S4), suggested that CSB-PGBD3 could potentially regulate gene expression without binding directly to DNA. To test if the CSB N-terminus alone can induce the changes in gene expression caused by CSB-PGBD3, we stably expressed two chimeric fusion proteins in UVSS1KO cells: a

CSB-LacI chimera in which the C-terminal PGBD3 domain is replaced by LacI, and the reciprocal eGFP-PGBD3 chimera in which the N-terminal CSB domain is replaced by eGFP. Using QPCR, we then compared the relative expression of selected genes in the stable lines expressing the CSB-LacI, eGFP-PGBD3, and control CSB-PGBD3 constructs. We selected a panel of 23 genes for the QPCR assay: 13 genes that were upregulated (signal log ratio SLR > 1) when CSB-PGBD3 was stably expressed in the CSB-null UVSS1KO line [26], 7 genes that were downregulated (SLR < -1), and 3 genes that showed no significant change in expression (SLR between 1 and -1).

Most of the 23 genes exhibited similar expression changes in both expression array and QPCR experiments: 14 genes were upregulated at least 2-fold (SLR>1) by CSB-PGBD3, 6 genes downregulated at least 2-fold (SLR<-1), and 3 genes exhibited less than 2-fold changes in expression by QPCR (Figure 3-15). The two exceptions were the v-src sarcoma viral oncogene homolog (SRC) which appears elevated by QPCR but not by microarray, and the spinocerebellar ataxia 1 gene (SCA1 or ataxin 1) which appeared to be less downregulated in the QPCR than in the microarray assay (SLR of -0.5 and -1, respectively). Thus the QPCR assays are consistent with our earlier expression array analysis [26].

Of the 14 genes upregulated by CSB-PGBD3, 11 were also upregulated by CSB-LacI, although less so for 8 of the 11 (Figure 3-15). Similarly, 8 of the same 14 genes were upregulated by eGFP-PGBD3, although less so for 6 of the 8 genes (Figure 3-15). These data suggest that both the N-terminal CSB domain and the C-terminal PGBD3 domain can independently upregulate genes induced by the CSB-PGBD3 fusion protein, but less effectively than when tethered together in a single protein. In contrast, the 6 genes downregulated by CSB-PGBD3 were almost unchanged by expression of CSB-LacI or eGFP-PGBD3 (Figure 3-15). CSB-LacI failed to downregulate any of these 6 genes by as much as the 2-fold cutoff for significance, and eGFP-PGBD3 downregulated only 1 of the 6 (Figure 3-15). Thus, downregulation of genes by CSB-PGBD3 requires fusion of the N- and C-terminal domains. Neither the CSB N-terminus nor C-terminal PGBD3 domain alone is capable of fully recreating the expression changes induced by CSB-PGBD3, and fusion of the two domains results in a transcriptional response that is greater than and somewhat different from the effect of the two domains individually. This could explain why the CSB-PGBD3 fusion has been conserved despite the presence of the CSB N-terminus in intact CSB and the intact PGBD3 protein transcribed from the cryptic promoter in CSB exon 5 (Figure 3-1 and [14]).

## Discussion

We have previously shown that expression of the CSB-PGBD3 fusion protein upregulates many genes related to innate immunity and an interferon-like antiviral response [26]. We also found that the PGBD3 domain of CSB-PGBD3 can bind MER85 elements *in vitro*, and therefore speculated that the CSB-PGBD3 fusion protein might regulate expression of nearby genes by binding to MER85 elements *in vivo*. Such binding could in principle affect gene expression in any of several ways: CSB-PGBD3 binding near many IFN-related genes or a few master regulators of the IFN response could drive an innate immune response directly. Alternatively, the N-terminal CSB domain of the CSB-PGBD3 fusion protein might act as a dominant negative in the absence of functional CSB, interfering with chromatin remodeling, and perhaps generating double-stranded RNA through bidirectional transcription, thus triggering innate immunity by mimicking a viral infection. And lastly, CSB-PGBD3 might affect gene expression not by binding site-specifically to MER85s but by interacting with unbound nucleoplasmic proteins. To begin to assess these models for regulation of gene expression by the CSB-PGBD3 fusion protein, we determined the genome-wide binding patterns of CSB-PGBD3, and used the Genomic Regions Enrichment of Annotations Tool (GREAT) to explore correlations between CSB-PGBD3 binding and gene regulation. As is often the case, the results were more interesting than the hypotheses.

Our genome-wide analysis of CSB-PGBD3 binding sites and related experiments have demonstrated that (1) CSB-PGBD3 is recruited not only to MER85 elements and MER85-related sequences within the PGBD3 transposon, but also to TRE, TEAD1, and CTCF motifs throughout the genome, as well as to sites of RNAPII enrichment in diverse cell lines; (2) binding of CSB-PGBD3 fusion protein to TRE motifs, but not to MER85s, correlates with genes upregulated by expression of CSB-PGBD3 in CSB-null UVSS1KO cells and genes downregulated by expression of functional CSB in CS1AN cells; (3) the CSB-PGBD3 fusion protein interacts with the TRE binding transcription factor c-Jun and RNAPII through the N-terminal CSB domain; (4) full regulation of genes by CSB-PGBD3 requires fusion of the CSB and PGBD3 domains; and (5) TRE motifs that bind CSB-PGBD3 are significantly enriched near bound but not unbound MER85 elements. These results suggest a far more complicated model for domestication of the CSB-PGBD3 fusion protein than we had originally anticipated (Figure 3-16). The CSB-PGBD3 fusion protein does indeed bind to MER85 elements throughout the genome as hypothesized, [R2] but these sites do not appear to correlate with regulation of nearby genes; instead, gene regulation reflects binding of CSB-PGBD3 to existing chromatin-bound transcription factors and, quite possibly, to RNAPII as well.

**A new layer of regulation on established regulatory networks**

Genome-wide binding of transposase-derived transcription factors had been demonstrated in Arabidopsis [120], but we provide a first look at the genome-wide binding of a transposase in transition: the PGBD3 transposase still binds strongly to related transposons, but has acquired novel functions because fusion with the N-terminus of CSB enables it to interact with previously established transcription factor networks. We had initially expected that binding of CSB-PGBD3 to MER85 elements would correlate with gene regulation induced by CSB-PGBD3. However, we found that CSB-PGBD3 interacts with a much broader range of binding sites, including TRE, TEAD1, and CTCF motifs, each of which is bound by factors that long predate horizontal transfer of PGBD3 to primate genomes. These results suggest that the conservation of the CSB-PGBD3 fusion protein over 43 My is due at least in part to modulation of existing regulatory networks rather than the creation of a *de novo* network based on insertion of MER85 elements near genes. However, the CSB-PGBD3 fusion protein also continues to bind MER85 elements, so we cannot rule out scenarios in which the fusion protein, bound to MER85 elements, regulates expression of nearby genes in specific cell types that we have not tested, or in occasional instances that would not appear statistically significant in our GREAT analysis. Thus a new protein (or RNA) that can modify established regulatory circuits may be able to build new functions without disrupting the old, whereas a new protein or regulatory RNA that can generate regulatory circuits de novo may be too powerful to survive because it would more likely do harm than good.

Nonetheless, transposable elements can, under very special circumstances, create regulatory networks *de novo*. For example, placental mammals express a large network of genes driven by transcription factor binding sites in MER20 transposons that regulate differentiation of endometrial stromal cells required for embryo implantation [22]. MER20s are present in >16,562 copies per human genome, 42% of which are located within 200 kb of the transcription start sites for pregnancy-induced genes. Moreover, the 218 bp element contains at least 22 potential binding sites for a total of 10 transcription factors (YY1, p300, C/EBP, CTCF, TGIF, p53, HoxA-11, FOXO1A, ETS1, and PGR), and quite remarkably 5 of these (C/EBP, PGR20, PGR21, FOXO1A, and HoxA11) are known to be important for hormone responsiveness and endometrial expression during pregnancy. The ability of MER20 to introduce a cluster of functional — and functionally related — transcription factor binding sites in a single insertional event may account for the evolutionary success of the MER20-based transcriptional network; it seems almost inconceivable that nearly identical clusters could have arisen at multiple genomic locations by neutral, stepwise mutation [121]. MER85s (some with bound CSB-PGBD3 fusion protein) could in principle participate in similar regulatory networks, as many MER85 elements contain binding sites for FOXA2, GFI, HAND1, HMGIY, HNF1A, NFE2L1, RORA, SOX5, and SRF (see Figure 3-17 for potential MER85 transcription

factor binding sites), but this seems less likely because MER85s are 20-fold less abundant than MER20s (889 versus 16,562 copies per genome).

**CTCF and CSB-PGBD3 may play roles in chromosomal looping**

CSB-PGBD3 may also regulate genes by affecting higher-order chromatin structure and looping. Our MEME analysis revealed the distinct signature of the CTCF binding motif in 58 CSB-PGBD3 peaks (Table S3, Figure 3-10). CTCF acts as a transcription activator or repressor depending on context, as a defining factor for gene insulation and silencing, and as a master regulator of long-range chromatin looping [110]. Although CTCF peaks represent only a small fraction of all CSB-PGBD3 binding sites, these peaks suggest that CSB-PGBD3 interacts directly with CTCF, perhaps mediating long-range interactions with CSB-PGBD3 bound to MER85s, TRE and TEAD1 motifs, or sites enriched for RNAPII. An interaction between CSB-PGBD3 and CTCF, through either the N-terminal CSB or C-terminal PGBD3 transposase domain, might also facilitate transposition. Intriguingly, the CTCF binding network in mammals has been shaped in part by retroposition of SINE elements that contain a CTCF motif [121], further expanding the repertoire of mechanisms by which transposons affect the structure and function of eukaryotic genomes.

CSB-PGBD3 could even play a direct role in chromosome looping (Figure 3-16). We found, by comparing all subsets of peaks in the ChIP-seq dataset for the CSB-PGBD3 fusion protein, that TRE motifs bound by the fusion protein are significantly enriched within 100 kb of MER85 elements that are also bound by fusion (see Results for details). Although it is possible that the fusion protein binds to these pairs of peaks independently, the data are consistent with chromosome looping mediated by the bifunctional fusion protein: the C-terminal PGBD3 transposase domain would bind to the MER85 and the N-terminal CSB domain would bind to AP-1 family transcription factors bound to the TRE motif — thus linking two distant sites, both of which would generate peaks in the ChIP-seq experiment.

**Additional roles for the N-terminal domain of CSB**

The full-length CSB protein plays an essential role in TC-NER by recognizing stalled RNAPII and initiating assembly of the large TC-NER complex [8,9,86]. As these interactions had not yet been mapped to specific domains of CSB, we were surprised to find that both the CSB-PGBD3 fusion protein and the chimeric CSB-eGFP protein are able to interact with RNAPII (Figure 3-14) although these proteins contain only the N-terminal domain of CSB and none of the 7 conserved ATPase motifs (Figure 3-1 and [26]). We do not yet know whether the interaction between RNAPII and the N-terminal domain of CSB occurs on DNA or at sites of stalled RNAPII, but our co-

immunoprecipitation experiments demonstrate that CSB and CSB-PGBD3 can share protein interaction partners through the common N-terminus. In fact, competition between CSB and CSB-PGBD3 for binding partners could play a role in CSB-dependent processes because expression of CSB-PGBD3 is about 4-fold higher than CSB in all cell lines we have examined [14].

The N-terminal domain of CSB has been shown to autoinhibit both normal and UV-induced association of CSB with chromatin [89,122], but deletion of the N-terminal acidic tract had no obvious effect on repair of UV damage [87,88]. Our data suggests that the N-terminus of CSB may play a larger role in targeting CSB to specific genes or chromosomal regions. We were surprised to find that the N-terminal domain of the CSB-PGBD3 fusion protein interacts both with both RNAPII (Figure 3-14) and c-Jun (Figure 3-12), and the sharp fragment accumulation profiles over TEAD1 and CTCF motifs (Figure 3-11) suggest that the fusion protein may also interact directly with TEAD1 and CTCF transcription factors bound to DNA. It also seems likely that the N-terminus of the CSB-PGBD3 fusion protein is responsible for binding to at least some of the 892 CSB-PGBD3 peaks (43% of 2,087 peaks total) that have no currently identifiable sequence motifs, but very likely bind transcription factors, chromosomal proteins, or enzymes such as topoisomerase I [37] involved in RNA and DNA transactions.

**A transcriptional role for CSB-PGBD3 in UV repair?**

UV irradiation and other stressors activate c-Jun through phosphorylation by c-Jun N-terminal kinases (JNKs, also called stress-activated kinases) such as JNK1 [123]. Activated JNK and AP-1 complexes can then affect cell proliferation and apoptosis, depending on cell type and stimulus [109,124]. We have previously shown that the CSB-PGBD3 fusion protein, although lacking all 7 ATPase motifs, can partially rescue UV damage repair in a host-cell reactivation assay using CSB-null UVSS1KO cells [26]. Conceivably, CSB-PGBD3 may facilitate repair by interacting with TC-NER proteins that normally associate with full-length CSB. However, the interaction between the CSB-PGBD3 fusion protein and the AP-1 family protein c-Jun (Figures 7, 9, S1) near genes upregulated by CSB-PGBD3 expression (Table S4) suggests an alternative scenario in which CSB-PGBD3 plays a transcriptional role in repair. After UV damage, activated AP-1 complexes could help guide CSB-PGBD3 to genes that are activated in response to UV. CSB-PGBD3 might then recruit RNAPII to these UV-activated TREs if the interactions of the N-terminal CSB domain of CSB-PGBD3 with c-Jun (Figures 7, 9, S1) and RNAPII (Figure 3-14) are not mutually exclusive.

**Does the CSB-PGBD3 fusion protein regulate CSB expression?**

We were surprised to find that PGBD3 is strongly bound by CSB-PGBD3 near three palindromic motifs that are also present in the 5' end of bound MER85s (Figure 3-5). Binding to these palindromes may autoregulate CSB transcription, CSB-PGBD3 expression, or alternative splicing and polyadenylation — perhaps by modulating the rate of RNAPII transcription or through interactions between the acidic N-terminus of CSB and phosphorylated serine/arginine-rich motifs in SR-family splicing enhancer proteins [125]. Thus it is possible that the CSB-PGBD3 fusion protein was initially retained in order to regulate CSB expression, and only secondarily acquired the ability to regulate other DNA repair, antiviral, and pathogen resistance genes.

**Do MER85s serve as a chromosomal reservoir for excess CSB-PGBD3 fusion protein?**

Continued binding of the CSB-PGBD3 fusion protein to MER85s may be fortuitous but need not be functionless. SETMAR (also called Metnase) is another domesticated transposase that exhibits continued binding to dispersed copies of the parental transposon. SETMAR consists of a SET methyltransferase domain fused to a Mariner (Hsmar1) transposase domain. SETMAR has been shown to play a role in NHEJ (nonhomologous end joining) repair of double-stranded DNA breaks [126] as well as repairing and restarting damaged replication forks [127], but it also retains the ability to bind Mariner transposon TIR sequences [128]. The binding affinity of SETMAR for Mariner elements appears to be regulated by interactions with a damage-regulated partner protein, Pso4 [129]; although normally bound to Mariner elements, SETMAR is released in response to DNA damage [130]. Similarly, MER85s could serve as reservoirs for excess CSB-PGBD3, perhaps regulating the interactions of CSB-PGBD3 with AP-1 factors, or ensuring that CSB-PGBD3 is readily available throughout the genome (Figure 3-16).

**The role of the CSB-PGBD3 fusion protein in Cockayne syndrome**

CS1AN cells are derived from a Cockayne syndrome patient with compound heterozygous CSB alleles. An early truncating mutation (K377term) in one CSB allele prevents expression of CSB and the CSB-PGBD3 fusion protein, but the 100 bp deletion in exon 13 of the other CSB allele [131] is located far downstream of PGBD3 and allows continued expression of the CSB-PGBD3 fusion protein in the absence of full-length CSB [14]. Surprisingly, genes downregulated by expression of full-length CSB in CS1AN cells [6] correlate strongly with CSB-PGBD3 binding sites in CSB-null UVSS1KO cells (Table S4). Thus, CSB-PGBD3 contributes to an aberrant transcriptional state in CS1AN cells by binding near, and perhaps interacting directly with, genes that are normally repressed by full-length CSB.

How could the N-terminal domain of CSB in the CSB-PGBD3 fusion protein activate genes that are normally repressed by CSB? One tantalizing but highly speculative scenario would be that for CSB-regulated genes, the autoinhibitory N-terminal domain of CSB [89] has dual checkpoint and transcriptional activation functions: Once the ATPase domain had engaged as a chromatin remodeler, the N-terminal domain would be released to activate transcription. Unconstrained in the fusion protein by the mutually autoinhibitory ATPase domain of CSB, the N-terminal CSB domain of CSB-PGBD3 would function as a constitutive transcriptional activator of CSB-regulated (and perhaps other) genes unless displaced by functional CSB.

GREAT analysis provided considerable insight into the consequences of the interactions of CSB-PGBD3 with TRE and TEAD1 motifs: Genes downregulated by CSB rescue of CS1AN cells correlate strikingly, for all regulatory domain sizes, with the entire set of CSB-PGBD3 binding sites including those with TRE, TEAD1, or no detectable motifs (green cells in Table S4). Thus, CSB-PGBD3 binding to each of these motifs, and even to the large number of peaks for which we could not identify a motif, correlates with upregulation of gene expression in CS1AN cells that continue to make the CSB-PGBD3 fusion protein but lack functional CSB. The correlation of CSB-PGBD3 binding sites with genes repressed by CSB in CS1AN cells suggests that the fusion protein substantially reshapes the transcriptome in CS patient CS1AN, and may do so in other CS patients whose mutations allow continued expression of the CSB-PGBD3 fusion protein in the absence of functional CSB.

Just as expression of functional CSB in CS1AN cells represses genes upregulated by continued expression of the CSB-PGBD3 fusion protein [6], so expression of the CSB-PGBD3 fusion protein in CSB-null UVSS1KO cells induces a strong interferon-related innate antiviral immune response which is dramatically repressed by coexpression of functional CSB [26]. This could be driven by CSB-PGBD3 binding to AP-1 binding motifs, which are known to play a role in upregulating pro-inflammatory cytokines [132] and chemokines such as IL-8 [133]. In normal aging, inflammation is driven by an increase in cytokine expression [134] and appears to be responsible for many age-related diseases [135]. Thus induction of AP-1 dependent inflammatory pathways by the CSB-PGD3 fusion protein may contribute to segmental aging in CS [136], and could be responsible for parts of the innate immune response (including IL-8) induced by CSB-PGBD3 expression in CSB-null UVSS1KO cells [26].

These observations suggest a previously unappreciated role for CSB in regulation of innate immunity and inflammation. Indeed, even CS patients who do not express the CSB-PGBD3 fusion protein because of mutations upstream of intron 5 (Figure 3-1) might inappropriately activate or fail to deactivate innate immune pathways. As

perceptively advocated by Brooks et al. [65], inflammation and calcification of the brain are seen both in CS and in another childhood neurodevelopmental disease known as Aicardi-Goutiéres syndrome (AGS). In AGS, loss of RNASEH2 or TREX1 nuclease activity causes accumulation of intracellular DNA and RNA fragments, counterfeiting a viral infection and triggering a constitutive type I interferon response [66]. Our data suggest that CS may also have an autoimmune component, caused both by loss of downregulation through CSB, and inappropriate upregulation by the CSB-PGBD3 fusion protein. If so, CS patients may benefit from treatment with immunosuppressive or anti-inflammatory drugs.

# Materials and Methods

### Identification of MER85 locations

We previously identified the locations of 613 partial or complete MER85 elements [14]. Closer examination of these elements revealed that almost all of them are actually complete, with both 5’ and 3’ terminal inverted repeats (TIRs). Additional MER85 elements in the hg18 build of the human genome were obtained from the RepeatMasker 3.2.7 track {RepeatMaskerOpen:tt} in the UCSC genome browser [117]. Several additional elements were located using the BLAT tool [137] in the UCSC genome browser with the 100 5’-most bases of the PGBD3 transposon as the query. Each MER85 element was examined individually to determine the boundaries of the sequence, the orientation, and the location of the TIRs and internal palindrome motifs (Table S1).

### Clones and Cell Lines

The HT1080 human fibrosarcoma cell line was maintained in Minimum Essential Medium Alpha (MEM-α) with 5% fetal bovine serum (FBS), penicillin, and streptomycin. All UVSS1KO-derived cell lines were cultured in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% FBS, penicillin, and streptomycin. UVSS1KO cells stably expressing the pFLAG-HA-CSB, pFLAG-HA-CSB-PGBD3 and pFLAG-HA constructs have been described previously [26]. To generate analogous pFLAG-HA-CSB-eGFP, pFLAG-HA-CSB-LacI, and pFLAG-HA-eGFP-PGBD3 constructs, the indicated coding sequences were fused in frame and inserted into the same bicistronic pIREShyg3 backbone (Clontech) (details available upon request; LacI was a gift of N. Maizels). The constructs were linearized before transfection into UVSS1KO cells (TransIT-LT1 transfection reagent, Mirus #MIR2300) and selection of stable pools with 200 µg/ml of hygromycin.

### Chromatin preparation

For ChIP-PCR, HT1080 cells were crosslinked with 1% formaldehyde for 10 minutes before quenching with 125 mM glycine. For ChIP-seq, UVSS1KO cells expressing FLAG-HA-CSB-PGBD3 were crosslinked with 0.5% formaldehyde for 5 min before quenching. Lower crosslinking was used for ChIP-seq to allow more thorough shearing of chromatin by sonication. Cells were then washed twice with phosphate buffered saline (PBS), scraped from the tissue culture plates, and resuspended in 1 ml cell lysis buffer (CLB, 5 mM PIPES pH 8, 85 mM KCl, 0.5% NP40) per 2 $10^7$ cells. Cells in lysis buffer were vortexed for 10 sec, incubated on ice for 10 min, and vortexed again for 10 sec. After lysis, nuclei were pelleted by centrifugation and resuspended in 500 µl RIPA buffer (10 mM TrisHCl pH 8, 140 mM NaCl, 1% Triton X-100, 0.1% SDS, 0.1% deoxycholate, 0.1 mM EDTA, 0.05 mM EGTA) per 2 $10^7$ cells. Glass beads (50 mg per 2 $10^7$ cells) were added to assist shearing, and nuclei were broken by 6 pulses of 10 sec each from a Sonic Dismembrator (Fisher Scientific) at a setting of 4 W. Chromatin samples were precleared by nutation for 1 h at 4°C with crosslinked Staph A cells (20 µl per 2 $10^7$ cells). Glass beads and Staph A were removed before immunoprecipitation.

**ChIP-PCR**

Rabbit polyclonal antibodies specific for the N-terminal 240 residues and the C-terminal 158 residues of CSB were raised against fusions with bacterial GST [78]. CSB antibodies were purified from GST antibodies by passage over a GST column [14]. Rabbit anti-mouse IgG, rabbit polyclonal anti-CSB-N-terminus, rabbit polyclonal anti-CSB-C-terminus, or no antibody was added to HT1080 chromatin preparations at a dilution of 1:200 and nutated overnight at 4°C. To precipitate bound antibodies, 0.1 vol Protein A-sepharose CL4B beads (Sigma) was added, and nutated for 1 h at 4°C. The beads were then washed 3x in RIPA buffer, 1x in RIPA wash buffer (10 mM Tris-HCl, pH 8, 250 mM LiCl, 0.1 mM EDTA, 0.5% NP40, 0.5% deoxycholate), and resuspended in 0.2 vol TE at pH 7.5. ChIP samples were digested with pancreatic ribonuclease A, followed by Proteinase K, and decrosslinked by incubation at 65°C overnight. ChIPs were assayed using PCR primers for the 6 genomic MER85 elements (Table S7) and α-$^{32}$P-dCTP to body-label the products.

**ChIP-seq library preparation**

ChIPs were performed as described for ChIP-PCR but using a 1:200 dilution of mouse monoclonal 1B1 directed against the N-terminus of CSB (kind gift of Hua-Ying Fan, University of Pennsylvania) and Protein G Dynabeads (Invitrogen) for the pulldown. An input sample of sheared, crosslinked chromatin was set aside from the same chromatin pool used for ChIPs. The input sample was digested with RNase, protease, and decrosslinked without enrichment by ChIP. The ends of the ChIP and input samples were repaired using End-It (Epicentre), A-tailed using Taq polymerase (Invitrogen), and Illumina paired-end sequencing adapters were ligated using Quick T4 DNA Ligase

(NEB). DNA fragments ranging from 400 to 700 bp were selected and purified by PAGE, then preamplified for 9 (input) or 12 cycles (ChIP) using Illumina paired-end preamplification primers, a BioRad iTaq supermix, and the following PCR protocol: denaturation 5 min at 95°C; cycling 30 sec at 95°C, 2 sec at 55°C, 2 sec at 72°C, and extension 10 sec at 72°C. For Illumina adapter and primer sequences, see supplementary methods of [100]. The preamplified samples were purified using a Qiagen PCR Cleanup kit, and were sequenced using an Illumina Genome Analyzer II (J. Shendure, University of Washington). Bases were called suing Illumina Real Time Analysis 1.5 software. Raw reads and processed data can be accessed at GEO study GSE37919.

**ChIP-seq read alignment**

The input sample generated 4,735,921 reads of 36 bp each. Reads from a single end of each sequenced fragment were aligned to the UCSC build hg18 (NCBI36) using the read alignment program Bowtie v0.12.7 with settings -n 0, -m 1, and --best to ensure that mapped reads had no mismatches, did not match multiple locations in the genome, and were from the best stratum of alignments [101]. All together, 3,307,313 reads were successfully aligned. The CSB-PGBD3 ChIP sample generated 14,263,776 paired-end reads of 36 bp each. These reads were aligned using the same settings as for the input, but were aligned as pairs with default settings for read spacing. All together, 8,574,668 paired-end reads were successfully aligned. Alignment files were created in both Bowtie and SAM format for subsequent analysis. After using IGVTools to sort the SAM-formatted files, the paired-end SAM-formatted Bowtie output was converted using a Perl script into fragment overlap WIG files (Appendix 3) for use with the Cis-regulatory Element Annotation System (CEAS) and for subsequent analysis of fragment overlaps. The Perl script used paired-end reads as boundaries for each sequenced fragment, looked for clusters of at least 5 overlapping fragments of a specified size, and then calculated the number of sequenced fragments overlapping each position of the genome within each fragment cluster (Appendix 3).

**ChIP-seq peak calling**

ChIP-seq peaks were called using three peak-calling programs. Model-Based Analysis of ChIP-seq (MACS) was used to call peaks using the full paired-end CSB-PGBD3 ChIP and Input datasets [102]. MACS identified 45,067 peaks with a p-value < 1e-5. For subsequent analysis, only the 9,835 peaks with a p-value <1e-12 were considered. ERANGE was also used for calling peaks after converting Bowtie map files to RDS format using the ERANGE makerdsfrombowtie Python script [103]. The ERANGE setting -minimum 2 was used to adjust the minimum enrichment threshold to 2-fold because of the disparate read depth between input and enriched samples. Using this setting, ERANGE found 3,743 peaks, which were used for subsequent analysis. The

third peak-calling algorithm used was Quantitative Enrichment of Sequence Tags (QuEST), with settings for transcription factor ChIP and custom peak calling parameters (20, 10, 3) [104]. QuEST identified 5,663 peaks, which were used for subsequent analysis. Peaks from each of the three peak-calling algorithms were compared using the Join on Genomic Intervals feature of the multi-purpose Galaxy analysis tool [138-140]. Comparison of peaks from all three algorithms showed 2,087 distinct enriched regions ("combined peaks") identified using all three algorithms. The 5' and 3' boundaries of these regions were determined based on the outermost boundaries of overlapping peaks identified using MACS, ERANGE, and QuEST. The summit of each combined peak was determined by using a Perl script to search for the center of the deepest fragment overlap in the CSB-PGBD3 WIG file generated as described above (Appendix 3). Combined peaks and peak summits are listed in (Table S2).

**Cumulative overlaps over CSB-PGBD3 binding elements**

Galaxy's Intersect tool was used to compare the locations of 2,087 Combined Peaks to the locations of the 889 MER85 elements (Table S1). Peaks were found to overlap 363 MER85 elements. The cumulative fragment overlap over all 363 bound MER85 elements was calculated using a Perl script (Appendix 3) to sum the fragment overlaps over each MER85 element in the whole-genome CSB-PGBD3 fragment overlap WIG file after correcting the positions of the fragments for MER85 position and orientation. The same script was used to generate fragment overlap profiles over individual PGBD3 and PGBD3 pseudogenes, as well as sets of all bound TRE, TEAD1, and CTCF motifs.

**Identification of non-MER85 motifs**

Combined peaks were filtered to remove peaks over MER85 elements using Galaxy's Subtract tool [138]. Using Galaxy's Extract Genomic DNA tool [138], 100 bp regions around each filtered peak summit (summit - 50 bp to summit + 49 bp) were extracted [138] and then searched using a local installation of the Multiple-Em for Motif Elicitation (MEME) tool with the settings -dna -mod zoops -minw 6 -maxw 12 -revcomp -nmotifs 5 [141]. The Position Specific Frequency Matrix (PSFM) for each of the 5 statistically significant motifs identified by MEME were submitted to the online version of Tomtom [108] for comparison to motifs in the JASPAR and TRANSFAC transcription factor motif databases. Of the 5 motifs, 3 matched known transcription factors AP-1, TEAD1, and CTCF. Using Galaxy's Subtract tool, 892 peaks were identified that contained none of the 3 transcription factor motifs. The summit sequences of these peaks were resubmitted to MEME to identify any additional motifs that may have been masked by the high-scoring AP-1, TEAD1, or CTCF motifs, but no additional motifs were identified.

**Correlation of ChIP-seq peaks with UVSS1KO and CS1AN microarray datasets**

Probe locations for all Affymetrix Human U133plus2.0 array probe sets were retrieved from the HG-U133_Plus_2 Annotations, CSV format, Release 31 (8/23/10) available on the Affymetrix web site. We had previously defined probe sets that were up- or downregulated at least 2-fold by expression of FLAG-HA-tagged CSB, CSB-PGBD3, or both compared to FLAG-HA tags alone in UVSS1KO cell lines [26], or by CSB rescue of the CS1AN cell line [6]. Lists of Affymetrix probes for each condition and the direction of expression change were compiled, and a Perl script used to convert each list of probes to a list of 5' ends of probe set locations for regulated probe sets. These "probe set start sites" were converted to regulatory domains using the GREAT createRegulatoryDomains program locally with settings for 1 Mb, 250 kb, or 100 kb maximum extensions [40]. The regulatory domains were then compared to sets of peaks using a Perl script that counted overlaps between the peak summits and the set of regulatory domains (script available on request). These counts were submitted to the GREAT calculateBinomialP program locally to obtain P-values (Table S4). False discovery rates (FDR) were then calculated empirically for each comparison using 100 sets of randomly selected summit locations of the same size as each peak set. P-values corresponding to a FDR below 1% were considered significant. Finally, the peak summits were compared to the regulatory domain tables that include gene names, and lists of genes with nearby CSB-PGBD3 peaks were generated for each comparison.

**Comparison of CSB-PGBD3 peaks to diverse gene ontologies**

Peak summit locations for all identified peaks, as well as for MER85, AP-1, TEAD1, CTCF, and other peaks separately, were submitted to GREAT v1.8.2 (great.stanford.edu) using default association rule settings for the hg18 genome build. The results from each analysis, including up to 20 significantly enriched gene sets for each database with a region enrichment of 2-fold or more and an FDR < 0.05, were downloaded as tab-separated files.

**Detection of AP-1 protein expression**

Whole cell lysates from UVSS1KO cell lines were separated by SDS-PAGE and western blotted as described previously [14]. Primary antibodies were rabbit polyclonal anti-JunD (sc-74, Santa Cruz Biotechnology), rabbit polyclonal anti-c-Jun (sc-1694, Santa Cruz Biotechnology), mouse monoclonal anti-Fra-1 (sc-28310, Santa Cruz Biotechnology), rabbit polyclonal anti-Fra-2 (sc-13017, Santa Cruz Biotechnology), mouse monoclonal ANTI-FLAG M2 (F3165, Sigma-Aldrich), and mouse monoclonal anti-actin (A2228, Sigma-Aldrich).

**Co-immunoprecipitations**

Subconfluent cells were trypsinized, washed in PBS, and counted. Nuclei were prepared by detergent lysis in CLB (1 ml/$10^7$ cells) and pelleted after 10 min on ice.

Whole cells or nuclei were resuspended in IP50 buffer [142] (10 mM TrisHCl, 50 mM KCl, 0.1 mM EDTA, 10% glycerol, 1 protease inhibitor cocktail [Roche]) at a concentration of $10^6$ nuclei/ml and sonicated for 10 sec using a Sonic Dismembrator (Fischer Scientific) at 4 W. Aliquots of 0.5 ml containing 5  $10^6$ cells or nuclei were nutated with a 1:200 dilution of antibody for 1 h, followed by a pulldown with Protein G Dynabeads (Invitrogen). The beads were washed 3 times with IP50 buffer, then resuspended in sample buffer, denatured, and resolved by 6% SDS PAGE (for RNAPII and FLAG-HA tagged proteins) or 10% SDS PAGE (for c-Jun). The gels were electroblotted to PVDF membranes, and blocked with 5% nonfat dry milk in Tris-buffered saline (TBS). Primary antibodies were added in blocking buffer containing 0.1% Tween 20, and the membrane washed 3 with TBS containing 0.1% Tween 20 (TBST). Horseradish Peroxidase (HRP)-coupled secondary antibodies were added in blocking buffer with Tween, then washed 3 with TBST. The HRP signal was detected on X-ray film using ECL Plus Western Blot Detection reagents (GE Healthcare).

**Quantitative PCR**

Total RNA was extracted from UVSS1KO-derived cell lines using TRIzol (Invitrogen). RNAs were reverse-transcribed using random primers (Invitrogen) and Superscript III (Invitrogen), then digested with pancreatic RNase A. cDNAs were purified on PCR Cleanup columns (Qiagen), quantified with a Nanodrop spectrophotometer, and used for QPCR. QPCR was performed using SYBR Green Master Mix (BioRad) with 1.25 μM primers and 20 ng cDNA per reaction. Primer sequences are available on request.

**Electrophoretic Mobility Shift Assays**

PGBD3 was expressed and MER85s were cloned as described in [26]. Electrophoretic mobility shift assays (EMSAs) were performed as described in [59]. Gels were dried, used to expose a storage phosphor screen, and scanned using a phosphorimager. Images were then quantitated using ImageJ [143], and the difference in intensity of shifted bands was compared between adjascent lanes from samples with and without addition of PGBD3 protein. These differences were then normalized by comparison with the scrambled sequence control (0%) and the MER85 consensus control (100%).

**Identification of transcription factor binding sites in MER85 elements**

Locations of MER85-39, MER85-236, MER85-592, and MER85-763 were converted to hg19 coordinates using the UCSC Genome Browser Convert tool, then were submitted to the online version of MAPPER2 [144] to locate transcription factor binding sites from the JASPAR database. Results were filtered to the $90^{th}$ percentile of MAPPER scores, then displayed in the UCSC Genome Browser using a feature of the MAPPER2 website. Transcription factor binding sites found in 3 of 4 MER85s were then used to search the Entrez Gene database [145] for human homologues.

# Figures

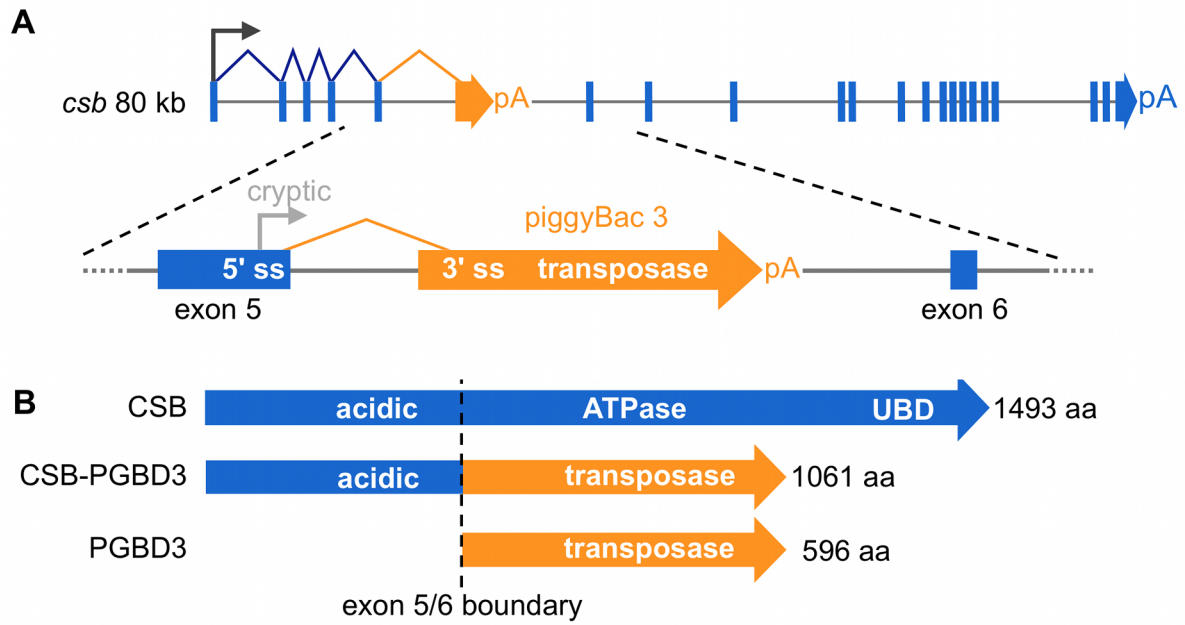



**Figure 3-1. The CSB-PGBD3 fusion protein is abundantly expressed by alternative splicing and polyadenylation of the CSB transcript.**
(A) The CSB-PGBD3 fusion protein is expressed by alternative splicing of CSB exons 1-5 to the PGBD3 transposase 3' splice acceptor site, whereas solitary PGBD3 transposase is expressed from a cryptic promoter in CSB exon 5. (B) As a result, the primate CSB locus generates three proteins: full-length CSB, the CSB-PGBD3 fusion protein, and solitary PGBD3 transposase. pA, polyadenylation signal; 5' ss, 5' splice donor site; 3'ss, 3' splice acceptor site.

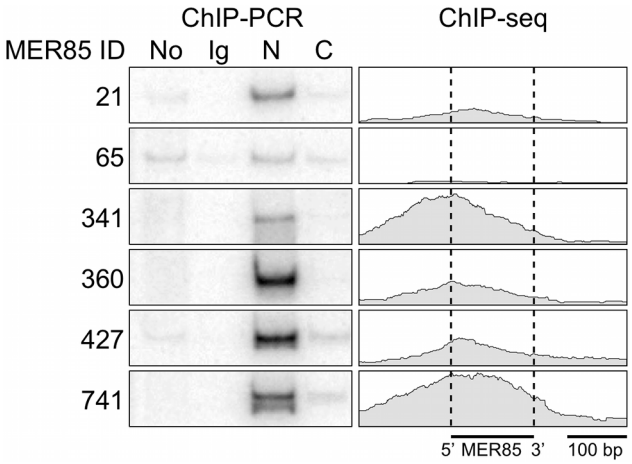



**Figure 3-2. The CSB-PGBD3 fusion protein binds to MER85 elements *in vivo*.** (*left*) ChIP-PCR in wild type HT1080 cells using antibodies for the N-terminus of CSB pulls down 6 representative MER85 elements with good matches to the Repbase consensus. N-terminal antibodies pull down both CSB-PGBD3 fusion protein and full-length CSB, whereas C-terminal antibodies pull down full-length CSB only (LTG unpublished). No, no antibody control; Ig, anti-mouse IgG nonspecific antibody control; N, CSB N-terminal antibody; C, CSB C-terminal antibody (*right*) Paired-end ChIP-seq shows enrichment for the same five out of six MER85 elements in CSB-null UVSS1KO cells stably expressing the CSB-PGBD3 fusion protein. Table S1 gives the positions and sequences of all MER85 elements. The 5' and 3' ends of MER85s are defined as the same orientation as the transposase ORF in parental PGBD3 elements.

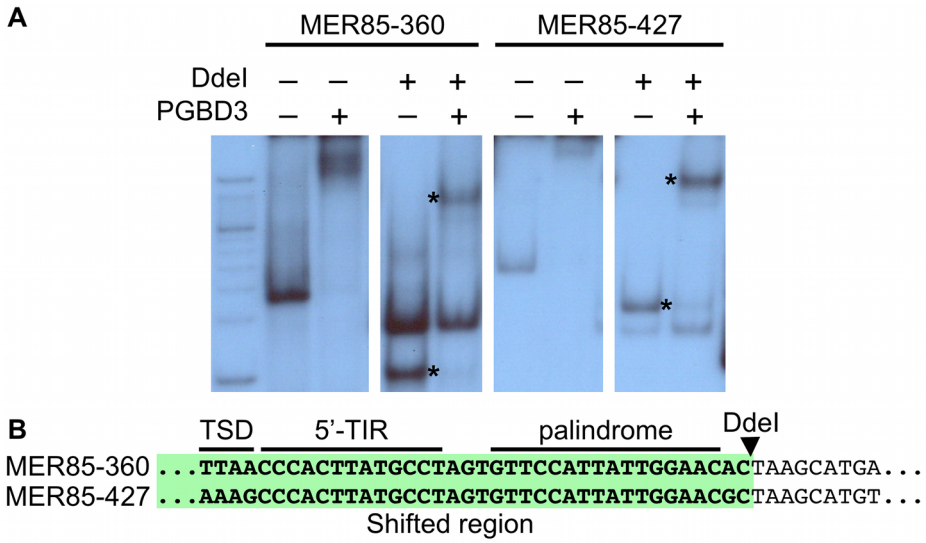



**Figure 3-3. The PGBD3 transposase binds to the 5' end of MER85s *in vitro*.** (A) An electrophoretic mobility shift assay (EMSA) using MER85s and MER85 DdeI restriction fragments. MER85-360 and MER85-427 were excised from plasmid clones using BamHI and EcoRI, then digested with DdeI or left intact. The restriction fragments were end-labeled and mixed with purified PGBD3 transposase. Restriction fragments derived from the 5' end of each MER85 are marked by an asterisk. (B) Partial sequences of MER85-360 and MER85-427 with the DdeI restriction site indicated. The 5' MER85 sequences that shifted upon incubation with transposase are highlighted in green. TSD, target site duplication; 5'-TIR, 5' terminal inverted repeat.

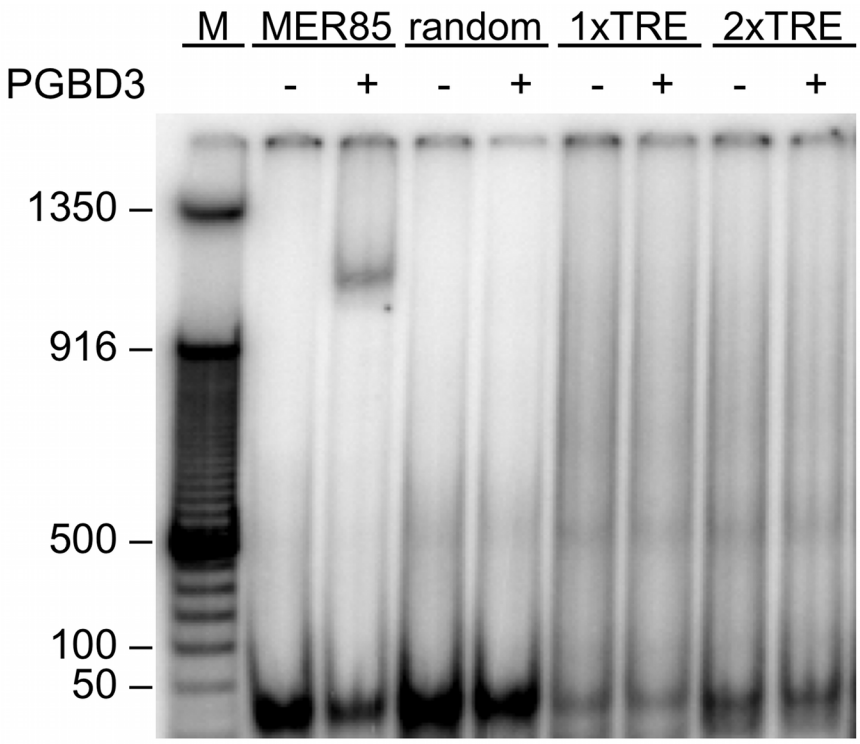



**Figure 3-4. The PGBD3 transposase is not capable of binding directly to TRE motifs *in vitro*.**
For EMSA assays, purified PGBD3 transposase was mixed with end-labeled 42 bp duplex oligonucleotides containing Repbase consensus MER85 sequence, 1 or 2 tumor promoting antigen response element (TRE) motifs, or random sequence.

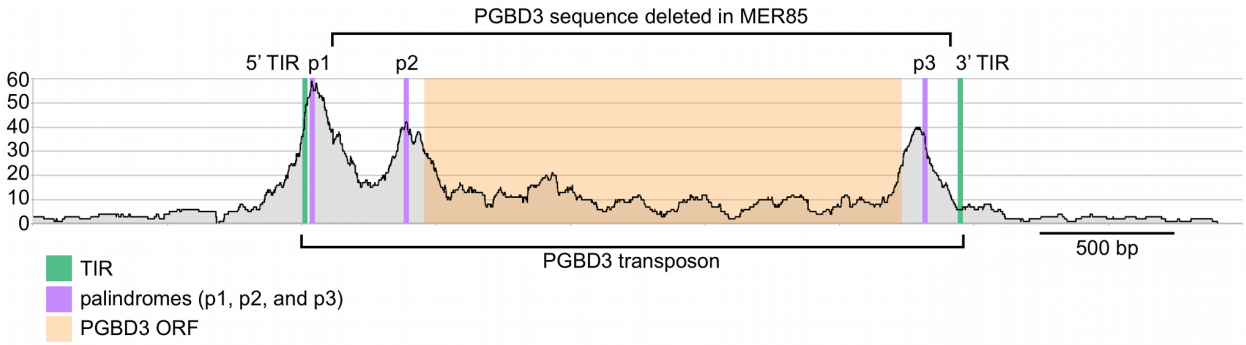
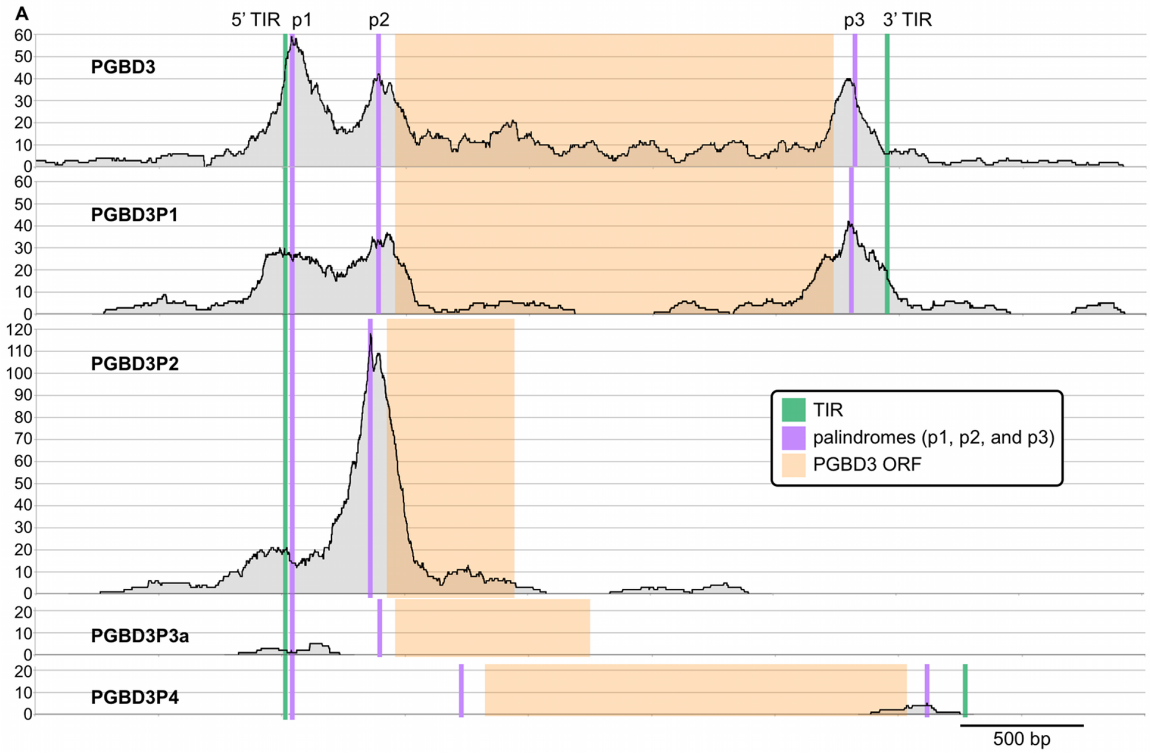



**Figure 3-5. Fragment overlaps in the vicinity of the PGBD3 transposon reveal strong binding near each of three palindromes.**
Occupancy of CSB-PGBD3 near the PGBD3 transposon was assessed by counting the number of overlapping ChIPed fragments at each position. TIR, terminal inverted repeat; ORF, open reading frame. Ordinate indicates the number of overlapping fragments.




B

| | | palindromes | | | |
|---|---|---|---|---|---|
| | 5' TIR | p1 | p2 | p3 | 3' TIR |
| PGBD3 | CC**T**ATTTATGCCT | GTTCCATTATTGGAAC | GTTCCATT**T**TTGGAAC | GTTCC**AA**TT**T**TGGAAC | AG**A**CATAAATGGG |
| PGBD3P1 | --CATTTATGCCT | GTTCCATTATTGGAAC | GTTCCATT**T**TTGGAAC | GTTCCATT**T**TTGGAAC | AG-CATAAA**CGA**G |
| PGBD3P2 | ------TATGCC**C** | GTTCCATTATTGGAAC | GTTCCATT**T**TTGGAAC | NP | NP |
| PGBD3P3 | ----TTTATACCT | GTTCCATTATTGGA**G**C | GTTCCATT**T**TT**A**GAAC | NP | NP |
| PGBD3P4 | ----**C**TTAT**G**CCT | GTTCCAT**C**ATTGGAA**A** | GTTC**T**ATT**T**TTGGAA**G** | GTTCCATT**T**TTT**T**GAA**T** | AGGCATAAA**A**GG**A** |

**Figure 3-6. Fragment overlaps over all full-length PGBD3 insertions in the genome, including all four PGBD3 pseudogenes, correlate with conserved palindrome sequences.**

(a) Fragment overlap binding profiles over PGBD3 and each of the PGBD3 pseudogenes. (b) Sequences of the TIR and palindromes of each of PGBD3 and each of the pseudogenes. Mismatches with respect to the PGBD3 p1 sequence are in bold. TIR, terminal inverted repeat; NP, sequence not present in truncated PGBD3 pseudogene.

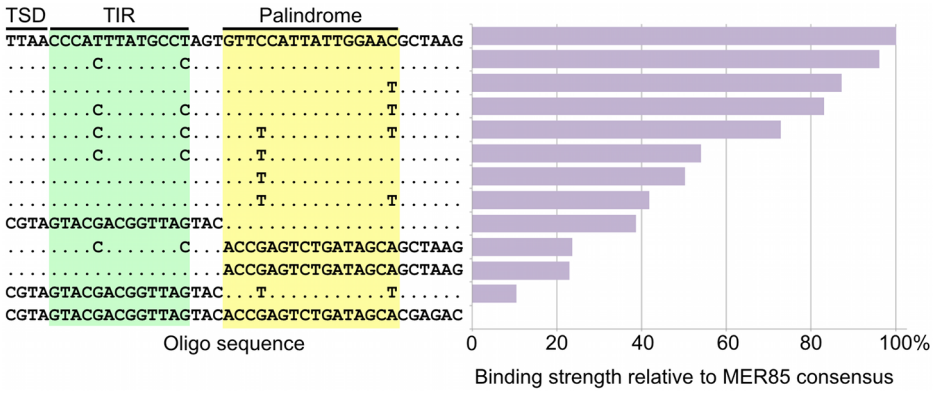



**Figure 3-7. Mutations in the palindromic region reduce PGBD3 transposase binding affinity for MER85s.**
Synthetic 42 bp MER85 fragments were mixed with purified PGBD3 or no protein, and used for an electrophoretic mobility shift assay. The binding affinities of the transposase for synthetic 42 bp fragments were normalized to the Repbase consensus sequence (100%) and a scrambled sequence (0%). Only sequence mismatches are displayed; positions that match the Repbase consensus are indicated by periods. TSD, target site duplication; TIR, terminal inverted repeat.

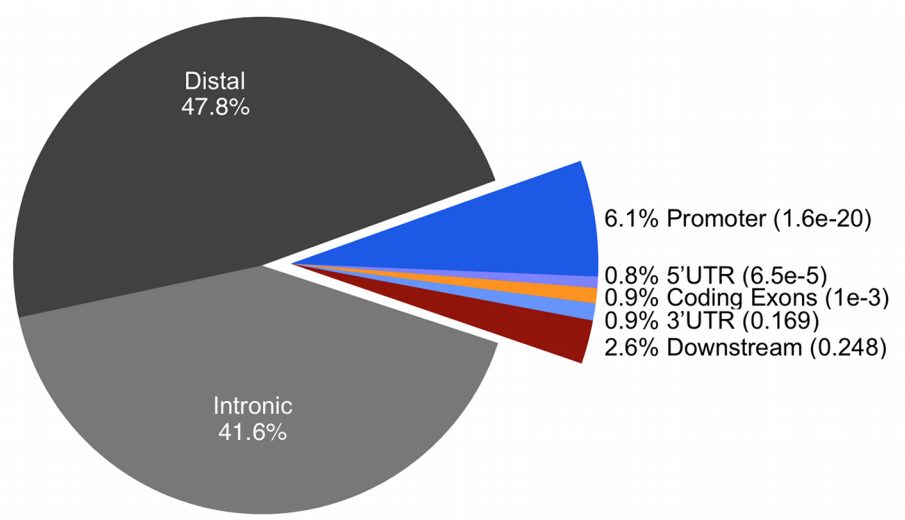



**Figure 3-8. CSB-PGBD3 fusion protein is enriched near gene promoters, but most peaks are distal and intronic.**
The Cis-regulatory Element Annotation (CEAS) Tool was used to generate a gene-centered annotation of 2,087 CSB-PGBD3 peaks found in common by MACS, ERANGE, and QuEST. Promoter regions include 3 kb upstream of the transcription start site (TSS). Downstream regions include 3 kb beyond the polyadenylation site. P-values generated by CEAS for overrepresentation of CSB-PGBD3 binding are shown in parentheses.

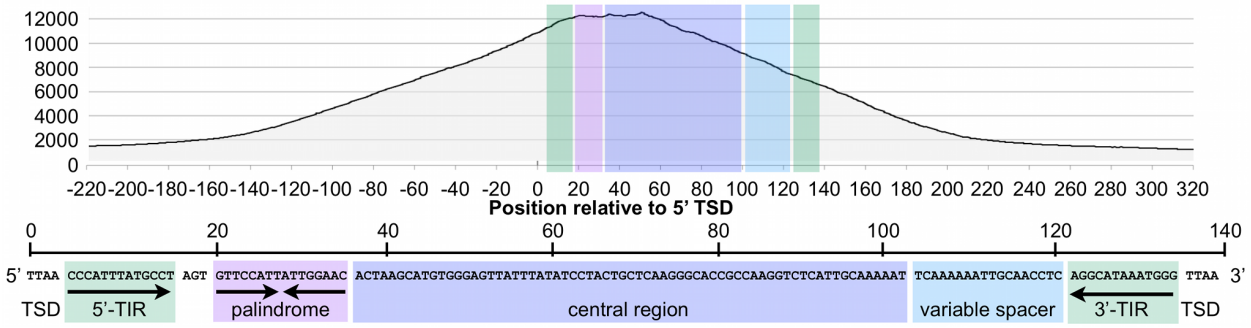



**Figure 3-9. The CSB-PGBD3 fusion protein binds preferentially to the 5' palindromic sequence of all bound MER85s in the human genome.**
Paired-end sequence reads near bound MER85s were used to reconstruct the location of immunoprecipitated fragments relative to the 5' target site duplication (TSD) of each element. Cumulative fragment overlaps were calculated by summing the number of fragments from each element that overlapped each position relative to the 5' TSD. TIR, Inverted Terminal Repeat.

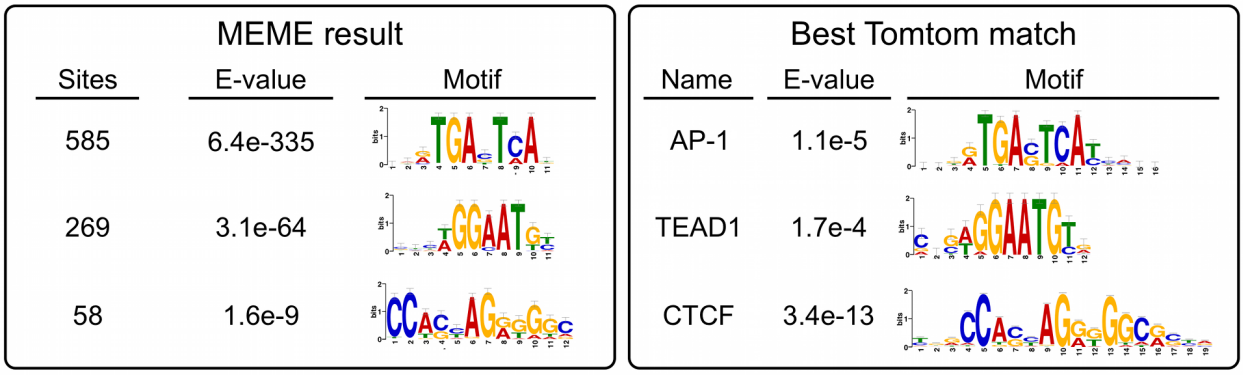



**Figure 3-10. Non-MER85 peaks are enriched for TRE, TEAD1, and CTCF binding site motifs.**
(*left*) Analysis using Multiple Em for Motif Elicitation (MEME). Sequences within 50 bp of non-MER85 peak summits were submitted to MEME to identify overrepresented motifs. (*right*) Analysis using Tomtom motif comparison tool. Position specific frequency matrices for the motifs identified by MEME were submitted to TOMTOM to identify matching transcription factor binding sites. The most significant matches for each result are shown. *AP-1 motif was annotated jundm2_secondary, Jun dimerization protein 2 secondary motif (UniPROBE mouse database); TEAD1, TEA domain family member 1 (JASPAR core 2009 database); CTCF, CCCTC binding factor (JASPAR core 2009 database).

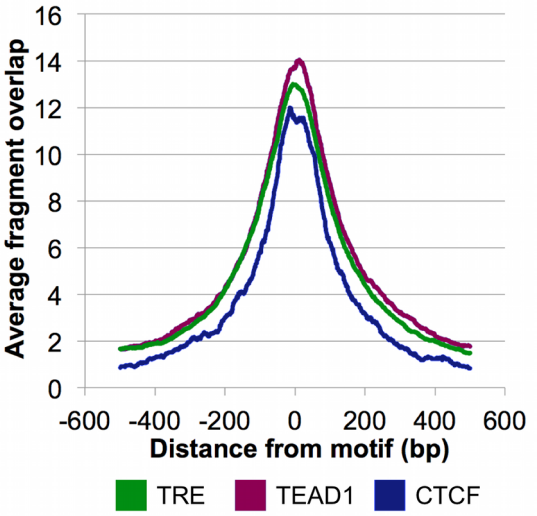



**Figure 3-11. CSB-PGBD3 peak summits coincide with the TRE, TEAD1, and CTCF motifs.**
Average fragment overlaps in the vicinity of TRE, TEAD1, and CTCF motifs were plotted for the CSB-PGBD3 ChIP-seq data. The overlaps peak sharply and symmetrically around the motifs, consistent with tethering of the CSB-PGBD3 fusion protein to the corresponding transcription factors through protein-protein interactions.

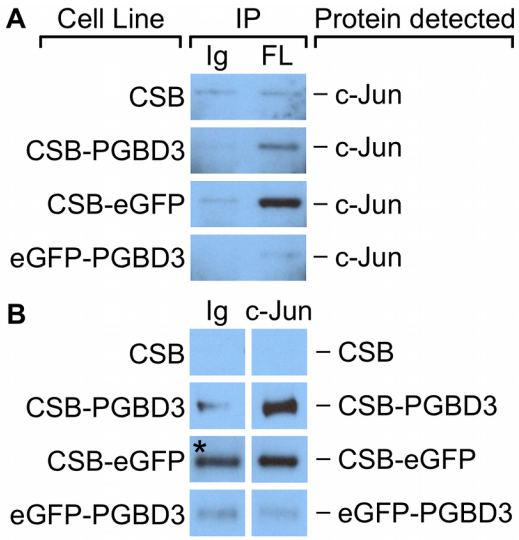



**Figure 3-12. c-Jun co-immunoprecipitates with the CSB-PGBD3 and CSB-eGFP proteins, but not with eGFP-PGBD3.**
Nuclear lysates from UVSS1KO cells stably expressing FLAG-HA-tagged CSB, CSB-PGBD3, CSB-eGFP, and eGFP-PGBD3 were immunoprecipitated using anti-FLAG, anti-c-Jun, and a nonspecific antibody. (*left*) Western blots probed with anti-c-Jun antibodies. c-Jun is immunoprecipitates with anti-FLAG antibodies in cells expressing FLAG-HA-tagged CSB-PGBD3 or CSB-eGFP, but not full-length CSB or eGFP-PGBD3. (*right*) Western blots probed with anti-FLAG antibodies. FLAG-HA-tagged CSB-PGBD3 and CSB-eGFP immunoprecipitate with anti-c-Jun antibodies. * Denotes lane with uncharacteristically high background. The same nonspecific antibody was used for all negative control samples, which leads us to believe this band is an artefact due to contamination rather than a true IP of CSB-eGFP. IP, antibodies used for immunoprecipitation; Ig, anti-mouse IgG nonspecific antibody control; FL, mouse monoclonal anti-FLAG antibody; c-Jun, anti-c-Jun antibody.

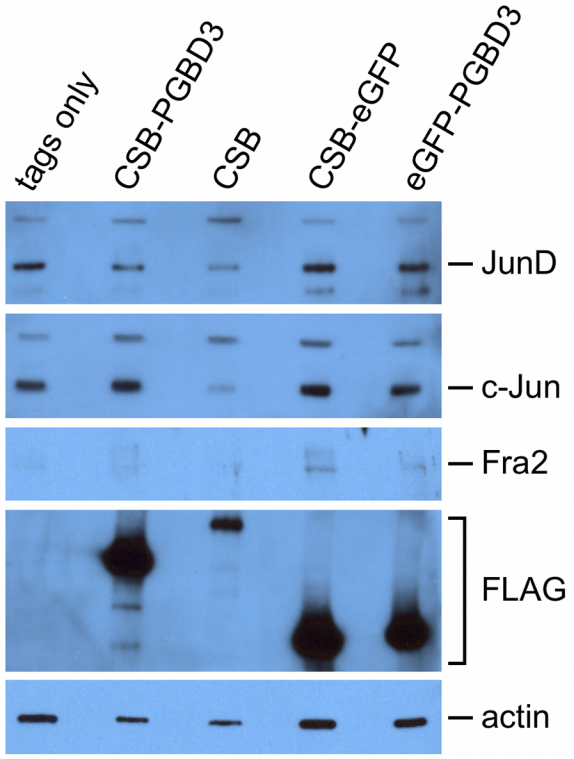



**Figure 3-13. c-Jun, JunD, and Fra2 are expressed in UVSS1KO cell lines.** Lysates from UVSS1KO cells expressing FLAG-HA tags or FLAG-HA-tagged CSB-PGBD3, CSB, CSB-eGFP, or eGFP-PGBD3 were western blotted for expression of JunD, c-Jun, Fra2, actin and FLAG-HA-tagged proteins.

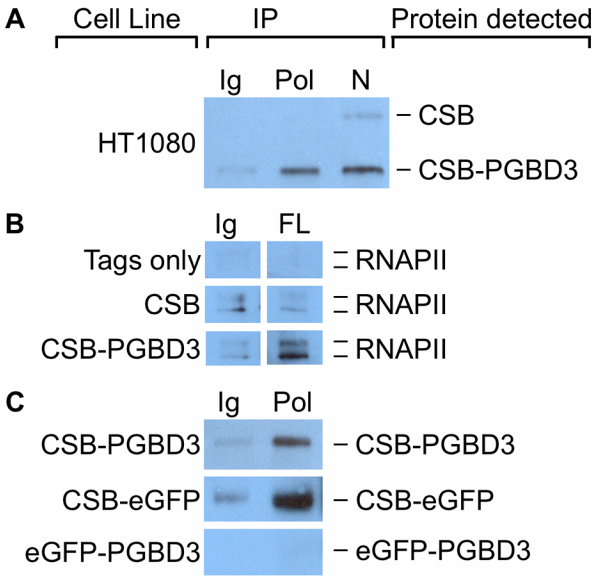



**Figure 3-14. CSB-PGBD3 and CSB-eGFP co-immunoprecipitate with RNA polymerase II (RNAPII).**
(A) HT1080 whole cell lysates were immunoprecipitated using anti-RNAPII CTD antibodies, N-terminal CSB antibodies, or nonspecific antibodies. CSB and CSB-PGBD3 were detected by western blotting with antibodies against the N-terminus of CSB. (B) UVSS1KO cells expressing FLAG-HA tags only, FLAG-HA-tagged CSB, or FLAG-HA-tagged CSB-PGBD3 were immunoprecipitated using antibodies for FLAG tags or a nonspecific antibody control. RNAPII was detected by western blotting with antibodies against the CTD of RNAPII. (C) UVSS1KO cells expressing FLAG-HA-tagged CSB-PGBD3, CSB-eGFP, or eGFP-PGBD3 were immunoprecipitated with antibodies against the CTD of RNAPII or a nonspecific antibody control. CSB-PGBD3, CSB-eGFP, and eGFP-PGBD3 were detected by western blotting with anti-FLAG antibodies. Ig, anti-mouse IgG nonspecific control; Pol, anti-RNAPII CTD; N, anti-CSB N-terminus; FL, anti-FLAG.

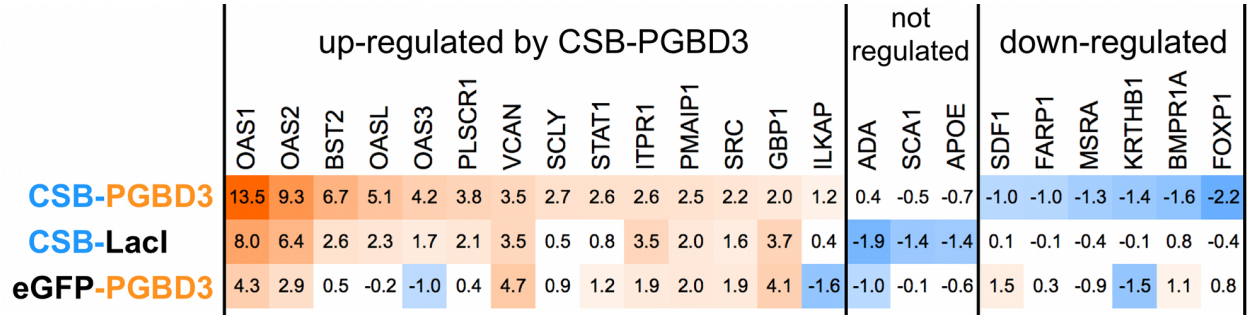



**Figure 3-15. CSB-LacI and eGFP-PGBD3 induce partial up-regulation of genes regulated by CSB-PGBD3.**
Average signal log ratios from quantitative real-time PCR (QPCR) of genes regulated by CSB-PGBD3, CSB-eGFP, or eGFP-PGBD3 expression in UVSS1KO cells compared to cells expressing FLAG-HA-tags alone. Orange cells: increased expression; Blue cells: decreased expression; No color: expression change was less than 2-fold (1 signal log ratio); darker color indicates a larger change in expression.

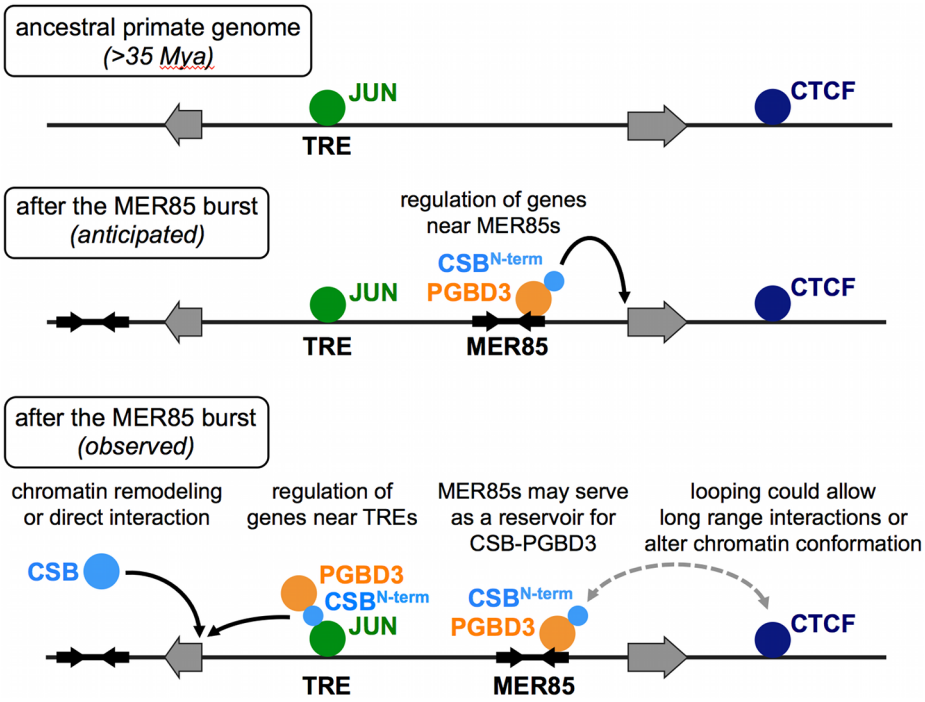



**Figure 3-16. ChIP-seq data suggest multiple roles for the CSB-PGBD3 fusion protein in gene regulation.**
*Top,*Transcription factor binding sites before the MER85 replicative burst. *Middle,* We anticipated that the CSB-PGBD3 fusion protein would bind to MER85 elements throughout the genome and regulate nearby genes through interactions mediated by the N-terminal CSB domain. *Bottom,* Our ChIP-seq data revealed that CSB-PGBD3 binds over TRE, CTCF, and TEAD motifs, and regulates genes near TRE motifs in CSB-null cells. Full-length CSB may facilitate or suppress these interactions through chromatin remodeling or competition for factors that also bind the N-terminal CSB domain of CSB-PGBD3. The CSB-PGBD3 fusion protein does bind to MER85 elements as anticipated, but these sites may function as a reservoir for CSB-PGBD3 protein or mediate chromatin looping, perhaps by interaction with CTCF.

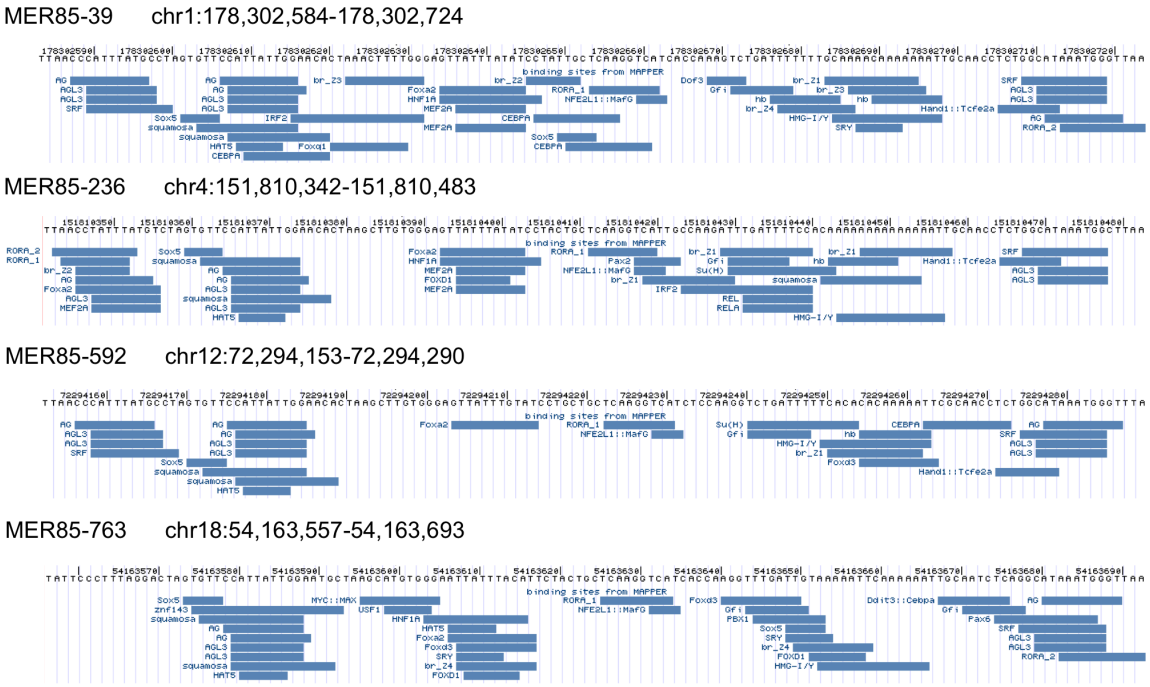



**Figure 3-17. MER85 elements contain potential transcription factor binding sites.** Locations of transcription factor binding sites from the JASPAR database found in 4 MER85 elements using MAPPER2 and displayed in the UCSC Genome Browser. MER85 sequences are shown from the upstream to downstream target site duplication (TSD).

| 1 Mb Regulatory Domains | | All Peaks (2087) | MER85 peaks (331) | TRE Peaks (585) | TEAD Peaks (269) | CTCF Peaks (58) | RNAPII Peaks (105) | Other peaks (803) |
|---|---|---|---|---|---|---|---|---|
| Cell Line | 2-fold direction | p-value | p-value | p-value | p-value | p-value | p-value | p-value |
| UVSS1KO + CSB | Down | **8.25E-05** | 8.64E-01 | **3.63E-03** | 9.33E-02 | 1.42E-03 | 1.62E-02 | 1.73E-02 |
| UVSS1KO + CSB | Up | 6.13E-02 | 2.22E-01 | 4.54E-02 | 4.46E-02 | 9.25E-02 | 4.71E-01 | 7.02E-01 |
| UVSS1KO + CSB-PGBD3 | Down | **1.90E-11** | 1.03E-01 | **1.39E-04** | 2.20E-03 | 1.16E-01 | 9.88E-02 | **1.07E-06** |
| UVSS1KO + CSB-PGBD3 | Up | **1.27E-08** | 1.62E-01 | **1.37E-05** | 8.53E-02 | 6.46E-03 | 6.38E-02 | **4.33E-04** |
| UVSS1KO + CSB + CSB-PGBD3 | Down | **1.41E-06** | 2.91E-01 | **4.10E-04** | 5.87E-01 | 3.25E-02 | 2.18E-02 | **5.16E-04** |
| UVSS1KO + CSB + CSB-PGBD3 | Up | **4.84E-15** | 7.06E-02 | **5.81E-06** | **4.30E-04** | 8.81E-03 | **8.70E-06** | **6.36E-06** |
| CS1AN CSB Rescue | Down | **2.82E-20** | 3.34E-01 | **2.73E-09** | **5.32E-06** | 1.97E-03 | **3.18E-03** | **8.64E-10** |
| CS1AN CSB Rescue | Up | 4.69E-02 | 5.38E-01 | 5.58E-02 | 9.84E-02 | 1.35E-01 | 1.25E-02 | 5.38E-01 |

| 250kb Regulatory Domains | | All Peaks (2087) | MER85 peaks (331) | TRE Peaks (585) | TEAD Peaks (269) | CTCF Peaks(58) | RNAPII Peaks (105) | Other peaks (803) |
|---|---|---|---|---|---|---|---|---|
| Cell Line | 2-fold direction | p-value | p-value | p-value | p-value | p-value | p-value | p-value |
| UVSS1KO + CSB | Down | 3.67E-02 | 8.51E-01 | 7.91E-02 | 3.74E-01 | 1.55E-01 | 1.29E-01 | 1.06E-01 |
| UVSS1KO + CSB | Up | 4.26E-01 | 2.59E-01 | 3.66E-01 | 3.49E-01 | 3.51E-01 | 1.00E+00 | 7.29E-01 |
| UVSS1KO + CSB-PGBD3 | Down | **2.90E-11** | 3.02E-02 | **3.47E-05** | **3.60E-04** | 1.10E-01 | 1.29E-01 | **3.90E-05** |
| UVSS1KO + CSB-PGBD3 | Up | **1.73E-05** | 2.33E-01 | **3.91E-05** | 6.11E-03 | 1.57E-02 | 4.08E-02 | 1.95E-01 |
| UVSS1KO + CSB + CSB-PGBD3 | Down | 1.35E-01 | 8.84E-01 | 1.20E-01 | 3.15E-01 | 1.29E-02 | 5.05E-01 | 3.39E-01 |
| UVSS1KO + CSB + CSB-PGBD3 | Up | **4.96E-13** | 1.83E-02 | **2.39E-04** | **1.40E-03** | 1.56E-02 | **9.48E-07** | **3.24E-04** |
| CS1AN CSB Rescue | Down | **1.20E-25** | 1.16E-01 | **1.69E-07** | **1.93E-07** | 4.68E-03 | 9.79E-03 | **3.27E-12** |
| CS1AN CSB Rescue | Up | 2.32E-01 | 6.05E-01 | 5.18E-02 | 1.76E-01 | 3.25E-01 | **6.24E-04** | 9.29E-01 |

| 100kb Regulatory Domains | | All Peaks (2087) | MER85 peaks (331) | TRE Peaks (585) | TEAD Peaks (269) | CTCF Peaks(58) | RNAPII Peaks (105) | Other peaks (803) |
|---|---|---|---|---|---|---|---|---|
| Cell Line | 2-fold direction | p-value | p-value | p-value | p-value | p-value | p-value | p-value |
| UVSS1KO + CSB | Down | 5.59E-01 | 6.50E-01 | 6.57E-01 | 9.14E-01 | 1.15E-02 | 5.29E-01 | 5.53E-01 |
| UVSS1KO + CSB | Up | 6.50E-01 | 4.01E-01 | 4.06E-01 | 2.12E-01 | 1.00E+00 | 1.00E+00 | 9.21E-01 |
| UVSS1KO + CSB-PGBD3 | Down | **1.32E-05** | 7.22E-02 | **6.21E-03** | 7.06E-02 | 5.43E-02 | 3.24E-01 | 1.74E-02 |
| UVSS1KO + CSB-PGBD3 | Up | 7.07E-03 | 1.03E-01 | **2.79E-03** | 1.53E-01 | 5.27E-03 | 3.94E-01 | 6.86E-01 |
| UVSS1KO + CSB + CSB-PGBD3 | Down | 2.16E-01 | 6.86E-01 | 4.51E-01 | 2.14E-01 | 7.65E-02 | 9.62E-01 | 3.01E-01 |
| UVSS1KO + CSB + CSB-PGBD3 | Up | **7.54E-06** | 5.65E-02 | **4.64E-03** | 9.57E-03 | 2.86E-01 | **3.65E-03** | 9.97E-02 |
| CS1AN CSB Rescue | Down | **1.37E-18** | 2.71E-02 | **1.76E-05** | **1.17E-06** | 5.89E-02 | **2.21E-03** | **1.81E-07** |
| CS1AN CSB Rescue | Up | 8.94E-01 | 8.16E-01 | 3.73E-01 | 1.94E-01 | 1.00E+00 | 6.29E-01 | 9.91E-01 |

**Table 3-1. Summary of GREAT comparisons to UVSS1KO and CS1AN expression arrays.**

Comparisons between CSB-PGBD3 binding sites and genes that exhibit expression changes of 2-fold or more when CSB, CSB-PGBD3, or both were stably expressed in CSB-null UVSS1KO cells, or when wild type CSB was stably expressed in the compound heterozygous CS1AN cell line which continues to express the CSB-PGBD3 fusion protein. The number of peaks in each set are given by numbers in parentheses. Bold p-values indicate a false discovery rate of less than 1%.

### All Peaks (2087)

| GO Biological Processes | P-value* | FDR** |
|---|---|---|
| positive regulation of catecholamine secretion | 4.4E-11 | 3.41% |
| positive regulation of amine transport | 2.6E-09 | 3.57% |
| positive regulation of secretion | 4.1E-09 | 2.57% |
| extracellular matrix organization | 6.7E-08 | 0.33% |
| positive regulation of cell adhesion | 1.0E-07 | 0.05% |
| **Disease Ontology** | | |
| malignant neoplasm of ovary | 6.0E-13 | 0.05% |
| cervix carcinoma | 8.3E-09 | 0.31% |
| Adenoviridae infectious disease | 2.6E-06 | 4.47% |
| pulmonary fibrosis | 3.4E-06 | 1.53% |
| diabetic retinopathy | 2.9E-04 | 1.42% |
| **Pathway Commons** | | |
| TNF receptor signaling pathway | 4.2E-15 | 0.01% |
| BMP receptor signaling | 1.8E-11 | 0.02% |
| Formation of Platelet plug | 6.1E-10 | 3.50% |
| IL2-mediated signaling events | 3.0E-09 | 0.41% |
| IL2 signaling events mediated by PI3K | 6.0E-09 | 2.72% |
| **MSigDB Perturbation** | | |
| Genes down-regulated in luminal-like breast cancer cell lines compared to the mesenchymal-like ones. | 1.9E-33 | 0.00% |
| Genes down-regulated in luminal-like breast cancer cell lines compared to the basal-like ones. | 8.7E-29 | 0.00% |
| Genes down-regulated in UB27 cells (osteosarcoma) at 12 hr after inducing the expression of a mutated form of WT1 [Gene ID=7490]. | 6.5E-23 | 0.01% |
| Genes down-regulated in primary fibroblast cell culture after infection with HCMV (AD169 strain) at 24 h time point that were not down-regulated at the previous time point, 20 h. | 3.0E-20 | 0.00% |
| Genes from common regions of gains observed in more than 15% of 148 primary breast cancer tumors. | 5.4E-17 | 1.38% |

### TRE Peaks (585)

| GO Biological Processes | P-value | FDR |
|---|---|---|
| blood vessel development | 3.5E-09 | 0.00% |
| vasculature development | 4.4E-09 | 0.00% |
| blood vessel morphogenesis | 9.5E-09 | 0.00% |
| angiogenesis | 1.2E-07 | 0.00% |
| response to oxygen levels | 1.7E-07 | 0.21% |
| **Disease Ontology** | | |
| respiratory system disease | 2.4E-09 | 0.01% |
| cervix carcinoma | 9.9E-08 | 0.01% |
| neurodegenerative disease | 1.4E-07 | 0.48% |
| melanoma | 4.3E-07 | 0.01% |
| melanocytic neoplasm | 5.3E-07 | 0.01% |
| **Pathway Commons** | | |
| IFN-gamma pathway | 5.8E-11 | 0.00% |
| Regulation of cytoplasmic and nuclear SMAD2/3 signaling | 1.2E-09 | 0.00% |
| TGF-beta receptor signaling | 1.2E-09 | 0.00% |
| Regulation of nuclear SMAD2/3 signaling | 1.2E-09 | 0.00% |
| ALK1 pathway | 3.0E-09 | 0.00% |
| **MSigDB Perturbation** | | |
| Genes down-regulated in luminal-like breast cancer cell lines compared to the basal-like ones. | 9.9E-14 | 0.00% |
| Genes down-regulated in luminal-like breast cancer cell lines compared to the mesenchymal-like ones. | 4.1E-13 | 0.00% |
| Genes up-regulated in U2OS cells (osteosarcoma) upon knockdown of HDAC1 [Gene ID=3065] by RNAi. | 3.6E-12 | 0.00% |
| Common down-regulated transcripts in fibroblasts expressing either XP/CS or TDD mutant forms of ERCC3 [Gene ID=2071], after UVC irradiation. | 7.3E-12 | 0.00% |
| Genes up-regulated in U2OS cells (osteosarcoma) upon knockdown of HDAC3 [Gene ID=8841] by RNAi. | 1.7E-09 | 0.00% |

### TEAD1 Peaks (269)

| GO Biological Processes | P-value | FDR |
|---|---|---|
| blood vessel morphogenesis | 9.3E-06 | 0.01% |
| angiogenesis | 4.9E-05 | 0.07% |
| blood vessel development | 5.3E-05 | 0.03% |
| **Disease Ontology** | | |
| malignant neoplasm of pancreas | 1.0E-07 | 0.00% |
| malignant neoplasm of female genital organ | 1.3E-07 | 0.26% |
| pancreatic neoplasm | 1.4E-07 | 0.00% |
| mature B-cell lymphocytic neoplasm | 1.8E-07 | 0.06% |
| female reproductive cancer | 1.9E-07 | 0.32% |
| **Pathway Commons** | | |
| IFN-gamma pathway | 2.4E-06 | 2.81% |
| ATF-2 transcription factor network | 1.7E-04 | 3.58% |
| Syndecan-2-mediated signaling events | 2.6E-04 | 3.87% |
| Noncanonical Wnt signaling pathway | 4.0E-04 | 3.36% |
| Canonical Wnt signaling pathway | 6.9E-04 | 4.72% |
| **MSigDB Perturbation** | | |
| Genes down-regulated in UB27 cells (osteosarcoma) at 12 hr after inducing the expression of a mutated form of WT1 [Gene ID=7490]. | 3.9E-08 | 0.09% |
| Genes down-regulated in UB27 cells (osteosarcoma) at 8 hr after inducing the expression of a mutant form of WT1 [Gene ID=7490]. | 5.9E-08 | 0.17% |
| HOXA9 [Gene ID=3205] targets down-regulated in hematopoietic stem cells. | 1.7E-07 | 2.28% |
| Genes down-regulated in luminal-like breast cancer cell lines compared to the mesenchymal-like ones. | 1.1E-06 | 0.00% |
| Genes associated with migration rate of 40 human bladder cancer cells. | 1.2E-06 | 0.01% |

### Peaks without motif (803)

| GO Biological Processes | P-value | FDR |
|---|---|---|
| positive regulation of defense response | 2.93E-05 | 4.30% |
| intracellular receptor mediated signaling pathway | 2.08E-04 | 0.89% |
| positive regulation of cell adhesion | 6.14E-04 | 0.39% |
| **Disease Ontology** | | |
| malignant neoplasm of ovary | 7.58E-07 | 4.37% |
| **Pathway Commons** | | |
| IL2 signaling events mediated by PI3K | 1.94E-08 | 0.33% |
| IL2-mediated signaling events | 9.53E-08 | 0.26% |
| BMP receptor signaling | 1.48E-07 | 0.22% |
| TNF receptor signaling pathway | 3.89E-07 | 0.19% |
| IL1-mediated signaling events | 4.08153E-06 | 0.38% |
| **MSigDB Perturbation** | | |
| Genes down-regulated in luminal-like breast cancer cell lines compared to the mesenchymal-like ones. | 3.36E-22 | 0.00% |
| Genes down-regulated in luminal-like breast cancer cell lines compared to the basal-like ones. | 2.68E-19 | 0.00% |
| Genes down-regulated in UB27 cells (osteosarcoma) at 12 hr after inducing the expression of a mutated form of WT1 [Gene ID=7490]. | 2.25E-16 | 0.00% |
| Genes whose expression positively correlated with sensitivity of breast cancer cell lines to dasatinib [PubChem=3062316]. | 1.65E-15 | 0.01% |
| Genes up-regulated in anaplastic thyroid carcinoma (ATC) compared to normal thyroid tissue. | 4.97E-15 | 0.00% |

**Table 3-2. GREAT results for comparisons of CSB-PGBD3 binding sites to diverse sets of gene ontologies.**

Only the five most significant results displayed by GREAT for each category are presented. Angiogenesis and blood vessel development genes are highly enriched near binding sites containing TRE or TEAD1 motifs. AP-1 proteins are known to regulate genes related to angiogenesis [92] although a role for TEAD1 in this process has not been studied. [R1] Genes involved in the related, but distinct transforming growth factor beta (TGF-beta) and bone morphogenesis protein (BMP) receptor signaling pathways are enriched near CSB-PGBD3 binding sites. The BMP receptor pathway is significantly enriched near the set of all CSB-PGBD3 peaks and the set with no identified motif, whereas TGF-beta receptor signaling is enriched near bound TRE motifs. SMAD2/3 signaling and the ALK1 pathway are also enriched near CSB-PGBD3 peaks over TRE motifs. All four of these factors – BMPR, TGF-beta, SMAD2/3, and ALK1 – are involved in overlapping pathways that regulate cell proliferation, bone growth, angiogenesis, and cell migration [146,147]. Disease Ontology and MSigDB perturbation terms related to [R1] breast, osteosarcoma, ovarian, cervical, melanocytic, and pancreatic cancers as well as GO Biological Processes for extracellular matrix organization and regulation of cell adhesion are enriched in one or multiple peak categories consistent with the notion that these ontologies reflect common pathways in oncogenesis. Immune response genes are enriched near CSB-PGBD3 binding sites, particularly those involved in IL-2 and IFN-gamma signaling. CSB-PGBD3 binding sites with TRE and TEAD1 motifs correlate strongly with IFN-gamma ontologies from Pathway Commons, [R1] while the set of all CSB-PGBD3 peaks and peaks with no identified motif were enriched near genes related to IL-2 signaling. GO, Gene Ontology; MSigDB, Molecular Signatures Database. *Raw binomial P-values reported by GREAT. **Hypergeometric false discovery rate (FDR) reported by GREAT.

## Supporting Table Legends

For the sake of brevity, the content of these tables is not included in the text of this document. They will be available as Supporting Materials of the online version of Gray, et al. (2012), PLoS Genetics 8, in press. The labels vary in the online version by omission of the "3-" prefix used here to denote that these tables are part of Chapter 3 of this thesis.

**Table 3-S1.** Locations and characteristics of PGBD3, 4 PGBD3 pseudogenes, and 889 MER85 elements in the hg18 genome. Sequences of the PGBD3 and MER85 elements, the 5' and 3' TIRs, and the palindrome are listed along with the peak ID from Table S2 for PGBD3s and MER85s that bound the CSB-PGBD3 fusion protein.

**Table 3-S2.** The CSB-PGBD3 fusion protein is enriched at > 2,000 locations in the hg18 human genome build. CSB-PGBD3 peaks were identified by all three peak finders (ERANGE, MACS, and QuEST) as significantly enriched by immunoprecipitation compared to the Input control. Summit locations were calculated as the region of greatest fragment overlap within the peak.

**Table 3-S3.** Genomic locations of TRE, TEAD1, and CTCF motifs identified by MEME within 50 bp of CSB-PGBD3 peak summits in the hg18 genome.

**Table 3-S6.** 105 CSB-PGBD3 peaks overlap at least 10 of 18 RNAPII peak datasets from 15 cell lines available in the UCSC Genome Browser database. The number of RNAPII datasets for which each CSB-PGBD3 peak overlaps an RNAPII peak are tabulated, and individual overlaps are annotated as 1 where an overlap was detected and 0 where no overlap was found. Peak IDs correspond to peak locations in Table S2.

**Table 3-S7.** Primers used for ChIP-PCR of genomic MER85 elements.

# Chapter 4:
# The PGBD3 gene may have been horizontally transferred between hydra and primates

## Summary

DNA transposons can be horizontally transferred between genomes, leading to the discontinuous representation of transposable elements in the genomes of related species or the appearance of an element in a single clade. In the case of PGBD3, we see both of these hallmarks of horizontal transfer. PGBD3 transposons are found in the genomes of Anthropoidea from marmoset to human, and in galago but not in lemurs or tarsier, which are thought to have diverged after the separation of Anthropoidea from galago. Thus, the presence of PGBD3 insertions, including the conserved CSB-PGBD3 fusion protein in the Anthropoidea, and the many decayed PGBD3 copies in galago, suggest that PGBD3 was horizontally transferred into the genomes of galago and the common ancestor of anthropoid primates on separate occasions. Here, we examine the sequences of all known PGBD3-like elements in primates, and present new evidence for horizontal transfer in the form of a surprisingly close homolog of PGBD3 found in the cnidarian *Hydra magnipapillata*. We show that several other piggyBac elements including human PGBD1 and PGBD2 function, like PGBD3, as alternative 3' exons resembling a gene trap vector to generate conserved, and presumably domesticated, mammalian piggyBac fusion proteins.

## Introduction

The CSB-PGBD3 fusion protein resulted from insertion of a piggyBac3 transposon into intron 5 of the CSB gene some 43 Mya. Although PGBD1 and PGBD2 are conserved in all mammals, and PGBD3 and PGBD4 are found only in anthropoid primates, the CSB-PGBD3 fusion protein is unique to anthropoid primates and extremely well conserved from marmosets to humans [14]. Thus, we suspect that PGBD3 was horizontally transferred into a common ancestor of all anthropoid primates, and that subsequent mutations within the PGBD3 transposase after insertion into the CSB gene may have facilitated exaptation of this element for a novel role in host gene regulation ([26] and Gray et al. (2012) PLoS Genetics 8, in press).

We have previously noted that there are dozens of PGBD3-like insertions in the genome of a non-simian primate, the Northern greater galago (*Otolemur garnettii*) [14], though all are now in a state of decay. Curiously, there appear to be no PGBD3 insertions in the mouse lemur genome despite the more recent divergence of lemurs and galagos ~57 million years ago (Mya) who themselves diverged from the common ancestor of simian primates ~77 Mya [148]. There are also no detectable PGBD3 insertions in the tarsier genome despite later divergence of tarsiers from simian primates ~65 Mya. Thus, the vector for PGBD3 transmission may have been active some time between ~57 Mya, when lemurs diverged from galago, and ~43 Mya, when the ancestor of humans diverged from marmosets (Figure 4-1).

We have recently found another PGBD3 homolog in a surprising place: the freshwater Cnidarian *Hydra magnipapillata* [149]. Hydra has been studied as a model organism for 300 years, and is known for its apparent lack of aging and ability to regenerate. The recently completed hydra genome contains a piggyBac transposase ORF that is far more closely related to the primate PGBD3 than any other protein in the NCBI protein database. This provides evidence for the horizontal transfer of PGBD3 between the exceedingly distant cnidarian and primate species. The direction of PGBD3 transfer, the age of the hydra insertion, and the steps required for horizontal transfer cannot be determined from currently available molecular information. Although transfer from hydra to primates might seem more likely, because an ancestral primate could have easily consumed hydra in fresh water, primate germlines appear to be safely sequestered from the gastrointestinal tract. In contrast, only a thin acellular layer of mesoglea separates germ cells in the hydra epidermis from the gastrodermis lining the gastrovascular cavity. These anatomical features may have provided a route for gene transfer from primates to hydra, and may explain the patchwork of mobile genes in hydra [149] including the broadly transferred *flp* gene [150]. In any event, we used hydra

PGBD3 as an outgroup to compare primate PGBD3s and to assess whatever subsequent mutations may have contributed to domestication of PGBD3 in primates.

Though there are several copies of PGBD3-like elements in the hydra genome, but only one appears to have a complete ORF. This hydra PGBD3 also appears to be expressed as an alternative 3' exon because the corresponding mRNA is annotated with an upstream exon that formally resembles the domesticated primate CSB-PGBD3 transcription unit. The ability of PGBD3 to function as a 3' splice acceptor in species as distant as hydra and human suggests that this is an important mechanism for host-independent expression of PGBD3 transposase: Gene trapping not only enables PGBD3 elements to achieve host-independent transcription by sharing host gene promoters but, because alternative splicing allows coexpression of the intact host gene protein along with the N-terminal transposase fusion protein, it enables PGBD3 to avoid the fate of other DNA transposons that function as insertional mutations. The unusually broad host range of piggyBac transposons [20], and the unusual ability of the piggyBac transposase from *Trichoplusia ni* (cabbage looper moth) to tolerate N-terminal fusions [151], provide further evidence that other piggyBac elements also use gene trapping as a survival strategy and shortcut to domestication. For example, the PGBD1 and PGBD2 elements are conserved among all mammals — PGBD1 as a fusion with an upstream zinc finger/SCAN domain, and PGBD2 as a fusion with two upstream non-coding exons.

## Results

### The hydra PGBD3 is more closely related to human PGBD3 than to any other known piggyBac transposon

A complete, active PGBD3 protein sequence would be ideal for analysis of the domestication of PGBD3 from transposase to transcription factor. Surprisingly, a protein BLAST search turned up a promising PGBD3 homolog in a most unexpected place: the recently sequenced genome of the freshwater cnidarian *Hydra magnipapillata*. Hydra are small, multicellular eukaryotes that have been studied as a model organism for 300 years [149]. A protein BLAST search showed that the hydra piggyBac is more similar to primate PGBD3 sequences than any other protein in the NCBI BLAST database (Table 4-1). Alignments of the hydra piggyBac to each of the human piggyBac-derived element proteins and the *Trichoplusia ni* piggyBac show that the hydra piggyBac is 66% identical to human PGBD3 (Figure 4-2 and Table 4-2). Human PGBD3 is also more similar to the hydra piggyBac than to any of the other human piggyBacs (Table 4-2). In contrast, the human and hydra enzymes GAPDH, hexokinase, DNA polymerase δ, and XPB averaged 48% identity; only the gene for the structural γ-tubulin protein was more highly

conserved at 85% identity. Such strong homology] between the PGBD3s of the very distantly diverged human and hydra genomes, as well as the absence of PGBD3 in any other metazoan genomes, suggests that this gene was horizontally transferred to hydra and primates from a common PGBD3 ancestor, or between hydra and the last common ancestor of Anthropoidea.

A BLAST search of the *H. magnipapillata* genome using the hydra PGBD3 DNA sequence shows that there are 12 additional locations in the genome assembly with homology to the PGBD3 transposon (Table 4-3). Aside from the full-length PGBD3 that was found initially, however, all of the other hydra PGBD3 copies are either pseudogenes with stop codons in the PGBD3 ORF, or have large truncations or internal deletions. The hydra PGBD3 sequence still contains all 3 catalytic aspartates in the transposase DDD/E motif, including D352 that is mutated to N352 in primate PGBD3 (Figure 4-2). However, the lack of multiple copies of highly similar PGBD3 transposons or MITEs (miniature inverted repeat elements [20]) suggests that the hydra PGBD3 is no longer an active transposase. Nonetheless, as the only non-primate PGBD3 sequence, the hydra PGBD3 can play an essential role in phylogenetic analysis of anthropoid PGBD3s. With hydra PGBD3 as an outgroup, we can more accurately compare primate PGBD3 sequences to determine the ancestral sequence of PGBD3 before domestication.

**Hydra PGBD3 is a 3' gene trap with the same 13 bp terminal inverted repeats (TIRs) as human PGBD3**

In addition to the protein sequence of the hydra PGBD3 transposon, we also examined the DNA sequence to see if the hydra transposon shares DNA features with the primate PGBD3s. We found that the hydra terminal inverted repeat (TIR) sequences are nearly identical to those of primate PGBD3. All piggyBac transposons have 13 bp recognition sequences, but the sequence of these repeats varies among piggyBacs. Thus, the matching sequence of both the hydra and primate PGBD3 TIRs again suggests that the hydra and primate PGBD3 share a common ancestor (Figure 4-3).

We have recently shown that the primary binding site of the human CSB-PGBD3 fusion protein to genomic MER85s is a palindromic sequence found 3 bp interior internal to the 5'-TIR of PGBD3 and MER85 elements. These palindromic sequences also flank the PGBD3 ORF, and were among the strongest binding sites for CSB-PGBD3 in the entire genome (L.T. Gray, K.K. Fong, T. Pavelitz, and A.M. Weiner (2012) PLoS Genetics 8, in press). These palindromic sequences are recognizable in the hydra PGBD3 transposon, but do not appear to be strongly conserved (Figure 4-3).

We also examined the 3’ splice acceptor site to see if the hydra PGBD3 could function as a 3’ terminal exon like the PGBD3 insertion within intron 5 of the ancestral anthropoid CSB gene. In both hydra and human, the AAG splice acceptor site is directly adjacent to the ATG start codon for the PGBD3 open reading frame (labeled TSS in Figure 4-3). Thus, the hydra PGBD3 could in principle be expressed in hydra as a 3’ gene trap. Indeed, consistent with this prediction, the hydra PGBD3 gene annotation notes an upstream exon that is spliced to the hydra PGBD3 ORF (Figure 4-4).

Finally, we found that both the human and hydra PGBD3 sequences have both a canonical polyadenylation signal (AATAAA) and a preferred polyadenylation site (CA) downstream of the PGBD3 translation termination codon (Figure 4-3). Though the position and spacing of these motifs differ, their presence in both PGBD3s suggests that both elements are capable of acting as 3' terminal exons, despite the computationally predicted annotation of another exon downstream of the hydra PGBD3 ORF (Figure 4-4).

**Using simian PGBD3 pseudogenes to reconstruct PGBD3’s past**

Having established that the hydra PGBD3 can serve as an outgroup, we assembled the sequences of PGBD3 pseudogenes (PGBD3P), which are present in 1 to 4 copies in each simian primate genome. These pseudogenes have decayed, and none encode a protein longer than 62 codons [14]. Assuming random decay, we can use these pseudogenes to reassemble the sequence of the active PGBD3 from which these pseudogenes arose. PGBD3P1 and PGBD3P4 are homologous to the entire PGBD3 sequence, while PGBDP2 has a 3’ truncation and PGBD3P3 a 3’ truncation and inversion. We obtained the sequences of PGBD3 and each PGBD3P in *Homo sapiens*, *Pan troglodytes* (chimpanzee), *Pongo abelii* (orangutan), and *Macaca mulatta* (rhesus macaque). Only two PGBD3P sequences (PGBD3P2 and PGBD3P3) were found in *Callithrix jacchus* (marmoset), and one (PGBD3P3) in *Papio hamadryas* (baboon), possibly because of gaps in the genome assemblies.

To remove insertions and deletions from the PGBD3 pseudogenes, we compared them to the conserved human PGBD3. Nucleotides ('N') were inserted or deleted in the DNA sequence to maintain the reading frame, then translated for phylogenetic analysis of protein sequences. Codons containing ‘N’ positions were translated as X and discarded for phylogenetic analysis. The inversion in PBD3P3 sequences was corrected to create the longest possible pseudogene sequence. The reconstructed pseudogene protein sequences were then aligned with human PGBD3 for use in phylogenetic analysis.

**Reassembling the galago PGBD3 sequence**

Several dozen PGBD3-like elements in the galago genome [14], all of which must have arisen after the divergence of galago and mouse lemur ~57 Mya (Figure 4-1), provided another valuable source of PGBD3 sequences. Most of these elements are in an advanced state of decay, and there appear to be no conserved copies. This suggests that PGBD3 was briefly and highly active in galago before all copies were silenced and subject to drift. This enabled us to reconstruct the original active galago PGBD3 sequence as the consensus of all existing partial PGBD3 sequences in the galago genome.

To obtain the PGBD3 sequences from galago, we used ENSEMBL BLAT against the current build of the galago genome (otoGar3) with the human PGBD3 transposon DNA sequence as bait. We found over 1000 statistically significant hits, most of which matched short sections of the PGBD3 sequence (Figure 4-5). To ensure that all PGBD3 hits were specific, we used only the 580 BLAT results that were at least 50 bp long. Many of these sequences came from adjacent sites in the galago genome, so regions within 500 bp of each other were joined to form 96 reconstructed PGBD3 elements. The sequences of these 96 reconstructed elements were aligned with each other using Clustal Omega, and Perl scripts were used to calculate the frequency of each base at each position in the alignment. Only those positions with at least 20 sequences reporting a base (to exclude those gapped by an insertion in one or more sequences) were used to calculate the most frequent base at each position of the collected galago sequences.

The consensus galago PGBD3 sequence was then aligned to the human PGBD3 transposon, and insertions and obvious frameshifts were removed so that the DNA sequence could be translated (Figure 4-6). After translation, we obtained a sequence that was 85% identical and 94% similar to the human PGBD3 sequence, although it still contained several stop codons near the 3’ end (Figure 4-7). These stop codons were removed before alignment to the human PGBD3 sequence for phylogenetic analysis.

**Phylogenetic analysis suggests the ancestral sequence of the primate PGBD3**

To determine what the sequence of PGBD3 may have been when it was still active, we used phylogenetic analysis to calculate the most likely protein sequence of the ancestral PGBD3 based on the anthropoid PGBD3 and PGBD3 pseudogene sequences, and the reconstructed galago PGBD3 sequence, using the more distant hydra PGBD3 sequence as an outgroup. Rather than simply take the consensus of aligned sequences, we used the PHYLogenetic Inference Package (PHYLIP) to infer the most likely ancestral protein sequence of PGBD3 using maximum likelihood analysis

(PROML). Each simian PGBD3 and pseudogene sequence (PGBD3P1-4) fell into distinct clades that closely resembled the phylogeny of simian primates. This shows that the insertions of PGBD3 and PGBD3P1-4 all occurred before the divergence of the common ancestor of simian primates. Surprisingly, the PGBD3 reconstructed from degenerate copies in galago appears most closely related to PGBD3P4 sequences, which suggests that PGBD3P4 is closer to the sequence of the horizontally transferred PGBD3 than the conserved PGBD3 ORF in intron 5 the CSB gene. PGBD3P2 appears to be the most distant of the PGBD3 pseudogenes, but the higher phylogenetic distance may simply reflect the consequences of 3' deletion] (Figure 4-8).

In addition to constructing phylogenetic trees, PHYLIP was used to calculate the most likely ancestral sequences for each internal node position in the tree. However, the ancestor of all available PGBD3 sequences cannot be calculated from this unrooted analysis. Instead, we compared the domesticated human PGBD3 sequence with the sequences of two internal nodes nearest to the hydra outgroup (Figure 4-8). In this comparison, we found that there were 31 positions in the amino acid sequences at which both of the calculated ancestral sequences differed from the human PGBD3 sequence, including the catalytic aspartate residue (D352 in both ancestral sequences, N352 in domesticated PGBD3). Of these, 23 were conservative changes, while 8 were nonconservative (including deletion of a glutamine residue at position 541 in the domesticated transposase).

**PSIPRED comparison to prokaryotic Tn5 transposase predicts PGBD3 transposase domain structure**

To help interpret the differences between conserved and ancestral PGBD3 sequences, we searched for similar proteins using the PSIPRED secondary structure prediction program and the pGenThreader fold search program. PSIPRED predicts the secondary structure of a protein sequence, and pGenThreader compares the predicted secondary structure to proteins with solved structures in the Protein Data Bank (PDB). A search using the PGBD3 protein sequence revealed that the secondary structure of PGBD3 transposase resembles that of the prokaryotic transposase Tn5 (Figure 4-9) which has been crystalized in complex with its target DNA [152]. Consistent with this structural prediction, the Tn5 transposase catalyzes DNA transposition through the same hairpin intermediate [152] as the *Trichoplusia ni* piggyBac transposase [116]. The secondary structure comparison between PGBD3 and Tn5 shows the location of the RNAseH-like transposase domain fold in the middle of PGBD3 (Figure 4-9). In Tn5, residues N-terminal to the catalytic domain are largely responsible for making key contacts with the terminal inverted repeats of the transposon sequence, some of which are distant from the catalytic site. The N-terminal regions of PGBD3 could also be important for this function, especially because it appears that PGBD3 interacts with both the TIR and

more distant palindromic sequences (L.T. Gray, K.K. Fong, T. Pavelitz, and A.M. Weiner (2012) PLoS Genetics 8, in press), perhaps simultaneously.

To ask if differences between the domesticated anthropoid PGBD3 and the reconstructed ancestral sequence might affect PGBD3 secondary structure, we submitted sequence PGBD3_A (Figure 4-8) to PSIPRED, and compared the predicted secondary structure of this “ancestral” PGBD3 sequence to that of the domesticated PGBD3 (Figure 4-10). No large differences were detected, which suggests that the overall structure of PGBD3 has been retained in the domesticated primate PGBD3, and the transposase fold may be important for the conserved function or stability of PGBD3.

**PGBD1 and PGBD2 are also conserved 3’ gene traps**

In addition to the PGBD3 genes in simian primates and hydra there are two other conserved piggyBac-derived genes that appear to be 3’ exon traps. PGBD1 and PGBD2 both have upstream exons that splice in frame with their piggyBac ORF, and appear to have been expressed as fusion proteins for at least 100 My (Figures 4-11c and 4-12c). PGBD1 has 6 upstream exons, 5 of which code for protein sequence that is fused to the long PGBD1 ORF. A protein BLAST search with these first 289 residues of the PGBD1 fusion protein revealed homology to zinc finger and SCAN domains (Table 4-4), many of which flank the PGBD1 fusion protein (Figure 4-11b and Table 4-4). Zinc finger gene clusters arise from expansion or tandem duplication events that occur in hotspots throughout the genome [153], and can be driven by increases in transposon activity [154]. Thus, even if insertion of the PGBD1 element disrupted host gene, nearby zinc finger gene duplications may have been functionally redundant and protected against loss of essential gene function. PGBD2 is also expressed and conserved as a 3’ splice acceptor (Figure 4-12), though the two exons upstream of the piggyBac ORF are mostly non-coding.

# Conclusion

Horizontal transfer of transposable elements has contributed both new proteins and new target sites for gene networks in eukaryotic genomes. Here, we have shown that a homolog of the domesticated primate PGBD3 transposase may also be conserved as a 3’ terminal exon in the cnidarian *Hydra magnipapillata*. We used this hydra PGBD3 as an outgroup to infer the sequence of the ancestral version of the conserved primate PGBD3, and to suggest that the secondary structure of primate PGBD3 has been conserved during domestication of the transposase. Finally, we show that PGBD1 and PGBD2 also survive as exon traps in mammals, and have been conserved for at least 100 My. Thus, the 3’ exon trapping lifestyle of these piggyBac transposons has poised

piggyBacs not only as useful insertional mutagens in the laboratory, but also for domestication as fusion genes in diverse genomes.

## Materials and Methods

### PSIPRED analysis of secondary structure

To calculate predicted secondary structures, we submitted protein sequences to the PSIPRED website (http://bioinf.cs.ucl.ac.uk/psipred/ and [155]). For human PGBD3, the Fold Recognition with pGenTHREADER option was selected to compare the predicted secondary structure to proteins in the Protein Data Bank.

### BLAT and BLAST searches

The locations of PGBD3 pseudogenes were obtained by BLAT searches [137] of primate genome builds on the UCSC Genome Browser website [117] using the human PGBD3 sequence as the query. The primate genome builds used were *Homo sapiens* build hg18, *Pan troglodytes* build panTro2, *Pongo abelii* build ponAbe2, *Macaca mulatta* build rheMac2, and *Callithrix jacchus* build calJac3. Sequences from the *Papio hamadryas* genome were obtained from BLAST of the preview build of the *P. hamadryas* genome available from the Pre!Ensembl database, and all of the galago (*Otolemur garnettii*) sequences were obtained by running BLAT on the Ensembl BLAST site (http://www.ensembl.org/Multi/blastview) [156]. The *Hydra magnipapillata* PGBD3 sequence was found using NCBI blastp with human PGBD3 protein sequence as input against the nonredundant protein sequence (nr) database [157,158]. The DNA sequence of the transposon was then retrieved using blastn against the Reference genome sequences (refseq_genomic) database [157,158].

### Sequence alignments

Alignment of protein sequences was performed using either ClustalW2 [159] or Clustal Omega [160]. Fine adjustment of sequence alignments performed prior to [PHYLIP analysis was carried out using SeaView [161].

### Phylogenetic analysis of aligned sequences

Phylogenetic analysis was performed using the PHYLIP 3.69 tool proml with default settings except for S (Not speedier, rough analysis) and 5 (Reconstruct hypothetical sequences) [162]. The dendrogram in Figure 4-8 was based on the distances provided as output from PHYLIP visualized using Dendroscope 3 [163].

### Genome browser screenshots

The genome browser view showing the upstream exon for the hydra PGBD3 was obtained from the NCBI GenBank entry for accession XM_002155126.1 [164]. All other genome browser views were obtained using the UCSC genome browser [117].

# Figures

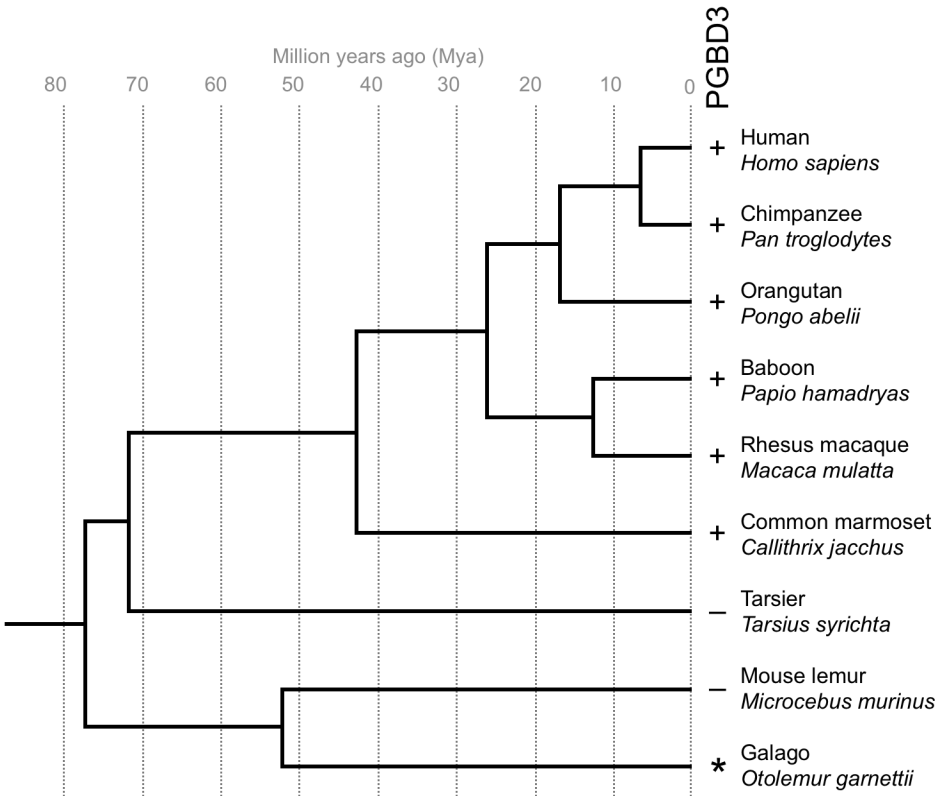



**Figure 4-1. Phylogenetic distribution of PGBD3 in primate species.**
A phylogenetic tree shows the discontinuous representation of PGBD3 in primate species. +, conserved PGBD3 ORF; –, no detectable PGBD3 insertions; *, only decayed PGBD3 insertions.

```
Hs_PGBD3        ----------MPRTLSLHEITDLLETDDSIEASAIVIQPPENATAPVSDEESGDEEGG  48
Hm_PGBD3        ----------MPKTLSFNEIMELLEEEDNIETSAIIIEPPVNANATNSDEDSGDEEGG  48
Hs_PGBD1        RECAPQIPCSTPIATERTVAHLNTLKDRHPGDLWARMHISSLEYAAGDITRKGRKKDKAR  60
Hs_PGBD2        RDVIAGRGIHSKVKSAKLLEVLNAMEEEESNNNREEIFIAPPDNAAGEFTDEDSGDEDSQ  60
Hs_PGBD4        ----------------MSNPRKRSIPMRDSNTGLEQLLAEDSFDESDFSEID--DSDNF  41
Tn_PGB          -------MGSSLDDEHILSALLQSDDELVGEDSDSEISDHVSEDDVQSDTEEAFIDEVHE  53
Hs_PGBD5        ----------MAEGGGGARRRAPALLEAARARYESLHISDDVFGESGPDSGGNPFYS  47

Hs_PGBD3        TINNLPGSLLHTAAYLIQDGSDAESDSDDPSYAPKDDSPDEVPSTFTVQQPPPSRRRKMT 108
Hm_PGBD3        TINNLPGSLLQAPAHLVND-SNFISNLDE----------EDVSISTVHELPKVKKIKLS  96
Hs_PGBD1        VSELLQGLSFSG-----DSDVEKDNE-PEIQPA-------QKKLKVSCFPEKSWTKRDIK 107
Hs_PGBD2        RGAHLPGSVLHASVLCEDSGTGEDNDDLELQPA-------KKRQKAVVKPQRIWTKRDIR 113
Hs_PGBD4        SDSALEADKIRPLSHLESDGKSSTSSDSG----------------RSMKWSARAMIP  82
Tn_PGB          VQPTSSGSEILDEQNVIEQPGSSLASNRILTLP-------QRTIRGKNKHCWSTSKSTRR 106
Hs_PGBD5        TSAASRSSSAASSDDEREPPGPPGAAPPPPRAP-------DAQEPEEDEAGAGWSAALRD 100

Hs_PGBD3        KILCKWKKADLTVQPVAGRVTAPPNDFFTVMRTPTEILELFLDDEVIELIVKYSNLYACS 168
Hm_PGBD3        KVTCKWKNKYLNIRPVKGRITPPQQ---TTAKTPTEIFELFLDDKVVEIIVTYTNLYASE 153
Hs_PGBD1        PNFPSWSALDSGLLNLK----------SEKLNPVELFELFFDDETFNLIVNETNNYASQ 156
Hs_PGBD2        PDFGSWTASDPHIEDLK----------SQELSPVGLFELFFDEGTINFIVNETNRYAWQ 162
Hs_PGBD4        RQRYDFTGTPGRKVDVS----------DITDPLQYFELFFTEELVSKITRETNAQAAL 130
Tn_PGB          SRVSALNIVRSQRGPTR---------MCRNIYDPLLCFKLFFTDEIISEIVKWTNAEISL 157
Hs_PGBD5        RPPPRFEDTGGPTRKMP----------PSASAVDFFQLFVPDNVLKNMVVQTNMYAKK 148

Hs_PGBD3        KGVH----------LGLTSSEFKCFLGIIFLSGYVSVPRRRMFWEQRTDVHNVLVSA 215
Hm_PGBD3        NGLT----------LGLSNSEFKFFLGIIILSGYVPLLRRRMFWEQRPDTHNKMVSS 200
Hs_PGBD1        KNVS----------LEVTVQEMRCVFGVLLLSGFMRHPRREMYWEVS-DTDQNLVRD 202
Hs_PGBD2        KNVN----------LSLTAQELKCVLGILILSGYISYPRRRMFWETSPDSHHHLVAD 209
Hs_PGBD4        LASKPPGPKGFSRMDKWKDTDNDELKVFFAVMLLQGIVQKPELEMFWSTRPLLDTPYLRQ 190
Tn_PGB          KRRESMTG------ATFRDTNEDEIYAFFGILVMT-AVRKDNHMSTDDLFDRSLSMVYVS 210
Hs_PGBD5        FQERFGSD------GAWVEVTLTEMKAFLGYMISTSISHCESVLSIWSGG-FYSNRSLAL 201

Hs_PGBD3        AMRRDRFETIFSNLHVADNANLDPVDK-----FSKLRPLISKLNERCMKFVPNETYFSFD 270
Hm_PGBD3        AMRRDQFSSIFTNLHVSNNANLDHEDK-----FAKIRPLISELNNRCLKFTPMETNFSFD 255
Hs_PGBD1        AIRRDRFELIFSNLHFADNGHLDQKDK-----FTKLRPLIKQMNKNFLLYAPLEEYYCFD 257
Hs_PGBD2        AIRRDRFELIFSYLHFADNNELDASDR-----FAKVRPLIIRMNCNFQKHAPLEEFYSFG 264
Hs_PGBD4        IMTGERFLLLFRCLHFVNNSSISAGQSKAQISLQKIKPVFDFLVNKFSTVYTPNRNIAVD 250
Tn_PGB          VMSRDRFDFLIRCLRMDDKSIRPTLREND--VFTPVRKIWDLFIHQCIQNYTPGAHLTID 268
Hs_PGBD5        VMSQARFEKILKYFHVVAFRSSQTTHG-----LYKVQPFLDSLQNSFDSAFRPSQTQVLH 256

Hs_PGBD3        EFMVPYFGRHGCKQFIRGKPIRFG--YKFWCGATCLGYICWFQPYQ--GKNPNTKHEEYG 326
Hm_PGBD3        ESMVPYFSRHGCKQFIRGKPIRFG--FKFWCGATPLGCICWFQPYQ--GKNPNTQFQEYG 311
Hs_PGBD1        KSMCECFD---SDQFLNGKPIRIG--YKIWCGTTTQGYLVWFEPYQEESTMKVDEDPDLG 312
Hs_PGBD2        ESMCEYFGHRGSKQLHRGKPVRLG--YKIWCGTTSRGYLVWFEPSQ--GTLFTKPDRSLD 320
Hs_PGBD4        ESLMLFKGPLAMKQYLPTKRVRFG--LKLYVLCESQSGYVWNALVHTGPGMNLKDSADGL 308
Tn_PGB          EQLLGFRGRCPFRMYIPNKPSKYG--IKILMMCDSGTKYMINGMPYLGRGTQTN---GVP 323
Hs_PGBD5        EPLIDEDPVFIATCTERELRKRKKRKFSLWVRQCSSTGFIIQIYVHLKEGGGPDGLDALK 316

Hs_PGBD3        VGASLVLQFSEALTEAHPGQYHFVFNN-FFTSIALLDKLSSM--GHQATGTVRKDHIDRV 383
Hm_PGBD3        VGASVVLQFVEALTEEHPGEYHFIFDN-FFTSISLLDKLTSM--GHHVTGTVRKDRIEKP 368
Hs_PGBD1        LGGNLVMNFADVLLERGQYPYHLCFDS-FFTSVKLLSALKKK--GVRATGTIRENRTEKC 369
Hs_PGBD2        LGGSMVIKFVDALQERGFLPYHIFFDK-VFTSVKLMSILRKK--GVKATGTVREYRTERC 377
Hs_PGBD4        KSSRIVLTLVNDLLGQG---YCVFLDN-FNISPMLFRELHQN--RTDAVGTARLNRKQ-- 360
Tn_PGB          LGEYYVKELSKPVHGSCR---NITCDN-WFTSIPLAKNLLQEPYKLTIVGTVRSNKREIP 379
Hs_PGBD5        NKPQLHSMVARSLCRNAAGKNYIIFTGPSITSLTLFEEFEKQ--GIYCCGLLRARKSDCT 374

Hs_PGBD3        PL---ESDVALKKKERGTFDYRIDG-KGNIVCRWNDNSVVTVASSGAGIHPLCLVSRYSQ 439
Hm_PGBD3        PL---DSDEVMKNKERGAFDYRSDD-KGNIICKWHDNSVVTVALSAAGINSIGLVNRYSQ 424
Hs_PGBD1        PL---MNVEHMKKMKRGYFDFRIEENNEIILCRWYGDGIISLCSNAVGIEPVNEVSCCDA 426
Hs_PGBD2        PL---KDPKELKKMKRGSFDYKVDESEEIIVCRWHDSSVVNICSNAVGIEPVRLTSRHSG 434
Hs_PGBD4        -----IPNDLKKRIAKGTTVARFCG--ELMALKWCDGKEVTMLSTFHNDTVIEVNNRNGK 413
Tn_PGB          EV---LKNSRSR--PVGTSMFCFDGPLTLVSYKPKPAKMVYLLSSCDEDASINESTG--- 431
Hs_PGBD5        GLPLSMLTNPATPPARGQYQIKMKG-NMSLICWYNKGHFRFLTNAYSPVQQGVIIKRKSG 433

Hs_PGBD3        KLKKKIQVQQPNMIKVYNQFMGGVDRADENIDKYRAS-IRGKKWYSSPLLFCFELVLQNA 498
Hm_PGBD3        KEKKRIQIQQSRMIKLYNQYMGGIDRADENIGQYQAS-ICGKKWYSSPLLFCFEVILQNA 483
Hs_PGBD1        DNEEIPQISQPSIVKVYDECKEGVAKMDQIISKYRVR-IRSKKWYSILVSYMIDVAMNNA 485
Hs_PGBD2        AAKTRTQVHQPSLVKLYQEKVGGVGRMDQNIAKYKVK-IRGMKWYSSFIGYVIDAALNNA 493
Hs_PGBD4        KTK------RPRVIVDYNENMGAVDSADQMLTSYPSERKRHKVWYKKFFHHLLHITVLNS 467
Tn_PGB          ---------KPQMVMYYNQTKGGVDTLDQMCSVMTCS-RKTNRWPMALLYGMINIACINS 481
Hs_PGBD5        EIP------CPLAVEAFAAHLSYICRYDDKYSKYFIS-HKPNKTWQQVFWFAISIAINNA 486

Hs_PGBD3        WQLHKTYDE-KPVDFL--EFRRRVVCHYLETHGHPPEPGQKGRP-QKRNIDSRYDGINHV 554
Hm_PGBD3        WLLHRTIEE-KPLDLL--EFHRRIASNYLETYGQPPETGRKGRASQKCNIDSRYDGLNHN 540
Hs_PGBD1        WQLHRACNPGASLDPL--DFRRFVAHFYLEHNAHLSD---------------------- 520
Hs_PGBD2        WQLHRICCQDAQVDLL--AFRRYIACVYLESNADTTSQGRRSRR---LETESRFDMIGHW 548
Hs_PGBD4        YILFKKDNPEHTMSHI--NFRLALIERMLEKHHKPGQQHLRGRPCSDDVTPLRLSGR--- 522
Tn_PGB          FIIYSHNVSSKGEKVQ--SRKKFMRNLYMSLTSSFMRKRLEAPTLKRYLRDN-ISNILPN 538
Hs_PGBD5        YILYKMSDAYHVKRYSRAQFGERLVRELLGLEDASP---------------------- 522

Hs_PGBD3        IVKQGKQTRCAECHKNTTFRCEKCDVALHVK-CSVEYHTE---------------- 593
Hm_PGBD3        IVNQNKQTRCRQCHKNTTFKCEKCSAALHIK-CSTKYHTHNNYIFFCDCIFL----- 591
Hs_PGBD1        ------------------------------------------------------- 
Hs_PGBD2        IIHQDKRTRCALCHSQTNTRCEKCQKGVHAK-CFREYHIR---------------- 587
Hs_PGBD4        --HFPKSIPATSGKQNPTGRCKICCSQYDKD-GKKIRKETR---------------- 560
Tn_PGB          EVPGTSDDSTEEPVMKKRTYCTYCPSKIRRK-ANASCKKCKKVICREHNIDMCQSCF 594
Hs_PGBD5        ----------------THYFCAECDVPLCVVPCFEIYHTKKNY-------------- 549
```

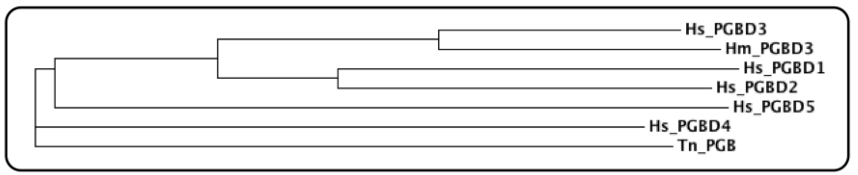


**Figure 4-2. The hydra piggyBac is most closely related to primate PGBD3**
Clustal Omega alignment of hydra PGBD3 with human PGBD3 and PGBD3Ps, PGBD1, PGBD2, PGBD4, PGBD5, the reassembled galago PGBD3, and *Trichoplusia ni* piggyBac. Bottom right panel, guide tree showing that Hs_PGBD3 and Hm_PGBD3 are more closely related to each other than to the other piggyBac sequences. Hs, *Homo sapiens*. Hm, *Hydra magnipapillata*. Tn, *Trichoplusia ni*.

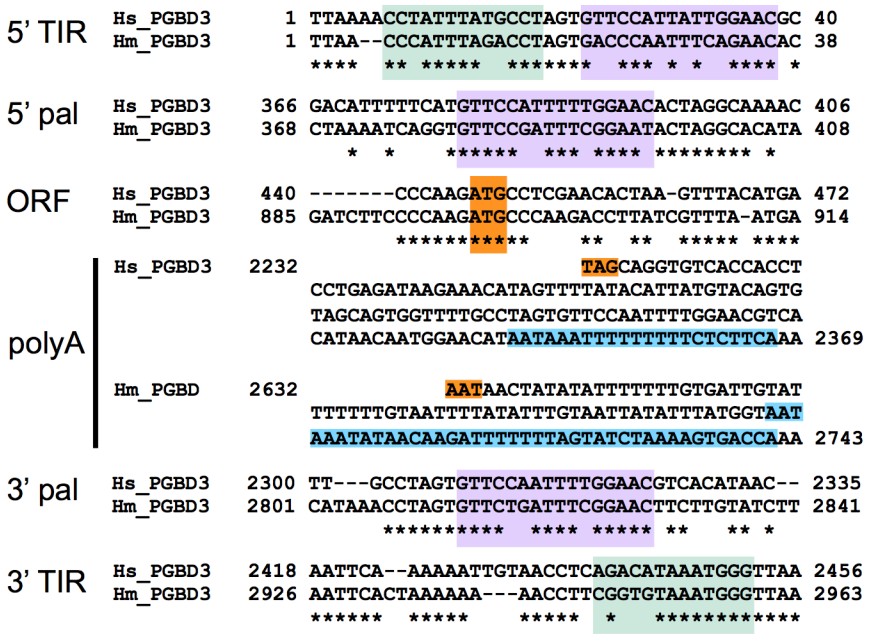



**Figure 4-3. The hydra and human PGBD3s have nearly identical TIRs, palindromic regions, and 3' splice acceptor sites.**
Alignments of selected regions of the human and hydra PGBD3 sequences show that the hydra PGBD3 has very similar TIR sequences (green). Some of the palindromic sequences in the hydra PGBD3 transposon are also very similar to PGBD3, though the 5'-most palindromic region is highly degenerate (purple). The 3' splice acceptor site (AAG) immediately precedes the transposase ORF in both species. Both human and hydra have polyadenylation signals (AATAAA) and potential polyA sites (CA), highlighted in blue, downstream of the end of the piggyBac ORFs (orange). 5' TIR, 5' terminal inverted repeat. 3' TIR, 3' terminal inverted repeat. 5' pal, 5' palindrome, found 34 bp upstream of the human PGBD3 ORF. ORF, open reading frame. 3' pal, 3' palindrome, found 74 bp downstream of the human PGBD3 stop codon. Hs, *Homo sapiens*. Hm, *Hydra magnipapillata*.

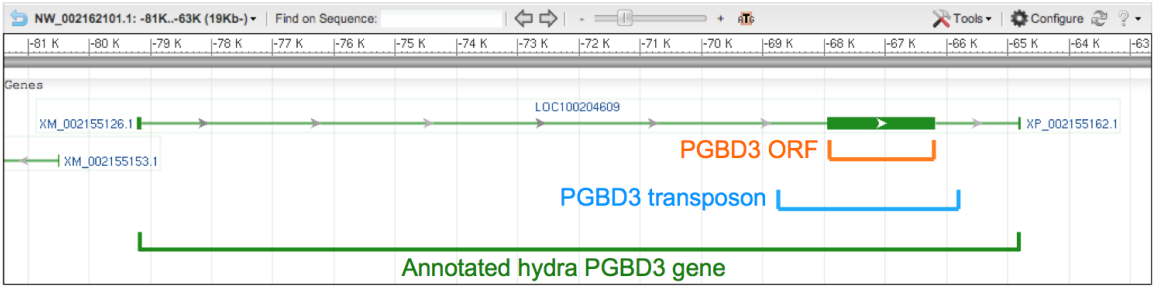



**Figure 4-4. The hydra PGBD3 gene has an upstream exon.**
The Genbank annotation for the hydra PGBD3 (XM_002155126.1) shows that the hydra PGBD3 ORF is predicted to be spliced to both a 5' and 3' exon. However, examination of the sequences flanking the PGBD3 ORF support splicing to a 5' exon, and polyadenylation at the 3' end of the ORF (Figure 4-3).


**Figure 4-5. BLAT for PGBD3 in the galago genome yields many hits.**
The alignment display generated by the ENSEMBL BLAT program shows the top 500 hits from the galago genome. These hits have good coverage over the ORF of the piggyBac transposase. HSPs, high scoring segment pairs above the BLAT default statistical threshhold. Only HSPs with homology of >50 bp were used for reconstruction of the galago sequences in Figure 4-6 and 4-7.

**Figure 4-6. DNA sequence alignment with PGBD3 used to eliminate insertions and frame shifts in the galago consensus PGBD3 sequence.**
Clustal Omega was used to align the DNA sequences of the galago and human PGBD3 elements. Red text denotes regions that were removed to maintain a consistent galago PGBD3 reading frame. Green highlight, beginning of PGBD3 ORF. Magenta highlight, PGBD3 ORF stop codon. Og, *Otolemur garnettii*. Hs, *Homo sapiens*.

```
Og_PGBD3     1 MPQTLSLHEITDLLETDDSIEASAIVIQPPENATAPVSDEDSGDEEGGTINNLPGSLLRT 60
Hs_PGBD3     1 MPRTLSLHEITDLLETDDSIEASAIVIQPPENATAPVSDEESGDEEGGTINNLPGSLLHT 60
               **:************************************:*****************:*

Og_PGBD3    61 PAYLIQDGYDAESDSDDPSYAPEDDSLDNEVPSTSTAQQPPPSKKKKVXKIVHKWKKADL 120
Hs_PGBD3    61 AAYLIQDGSDAESDSDDPSYAPKDDSPD-EVPSTFTVQQPPPSRRRKMTKILCKWKKADL 119
                ******* *************:*** * ***** *.******:::*: **: *******

Og_PGBD3   121 TAQPIAGRVTEPPNDFFTKMRTPTEILELFLDDEVVELIVMYSNLYAASKGVNLGLTSSE 180
Hs_PGBD3   120 TVQPVAGRVTAPPNDFFTVMRTPTEILELFLDDEVIELIVKYSNLYACSKGVHLGLTSSE 179
               *.**:***** ******* ****************:**** ******.****.*******

Og_PGBD3   181 FKCFLGIIFLSGYXSVPRRRMFWEQRTDVHNVLVSAAMRRDHFETIFSNLHVADNANLDP 240
Hs_PGBD3   180 FKCFLGIIFLSGYVSVPRRRMFWEQRTDVHNVLVSAAMRRDRFETIFSNLHVADNANLDP 239
               ************* ***************************:******************

Og_PGBD3   241 MDKFSKLQHLISTLNERCMKFFPNETYFSFDECMVPYFGHHGCKQFIRGKPIXFGCKFWC 300
Hs_PGBD3   240 VDKFSKLRPLISKLNERCMKFVPNETYFSFDEFMVPYFGRHGCKQFIRGKPIRFGYKFWC 299
               :******: ***.********.********** ******:************ ** ****

Og_PGBD3   301 GATRLGYISWFQPYQGKNPNTKHEKCGVGASLVLQFSEALTEAHPGQYHFVFXNFFTSIA 360
Hs_PGBD3   300 GATCLGYICWFQPYQGKNPNTKHEEYGVGASLVLQFSEALTEAHPGQYHFVFNNFFTSIA 359
               *** ****.***************: ************************** *******

Og_PGBD3   361 LLDKLSSMGHQATGTVRKDHIDKAPLESDVALKKKERGTXNYQIDGKGNIVCRWNDNSVV 420
Hs_PGBD3   360 LLDKLSSMGHQATGTVRKDHIDRVPLESDVALKKKERGTFDYRIDGKGNIVCRWNDNSVV 419
               **********************:.*************** :*:******************

Og_PGBD3   421 TVASSGAGIDPLCLVNRYSQKLKKKIQVQQPNMIKVYNHFMGGIDRADDNIDKYXASIHG 480
Hs_PGBD3   420 TVASSGAGIHPLCLVSRYSQKLKKKIQVQQPNMIKVYNQFMGGVDRADENIDKYRASIRG 479
               *********.*****.**********************:****:****:***** ***:*

Og_PGBD3   481 KKWYSSHSSPLLFCFKLVLQNAWQLHKTYDEKPVDFLEFH*RVVCHYLETHGHPPEPG*R 540
Hs_PGBD3   480 KKWYS---SPLLFCFELVLQNAWQLHKTYDEKPVDFLEFRRRVVCHYLETHGHPPEPGQK 536
               *****   *******:***********************: ***************** :

Og_PGBD3   541 GRPSQKHNIDSCYDGMNHVIVKQGKQMXCAECHKNTTF*CEKCDVALHVKCSVEYHTE 597
Hs_PGBD3   537 GRP-QKRNIDSRYDGINHVIVKQGKQTRCAECHKNTTFRCEKCDVALHVKCSVEYHTE 593
               *** **:**** ***:**********  ********** ******************
```

**Figure 4-7. The reconstructed galago PGBD3 sequence is very similar to the human PGBD3 in CSB.**
A Clustal Omega alignment of the reconstructed PGBD3 from the consensus of 96 galago PGBD3 fragments. Og, *Otolemur garnettii*. Hs, *Homo sapiens*.

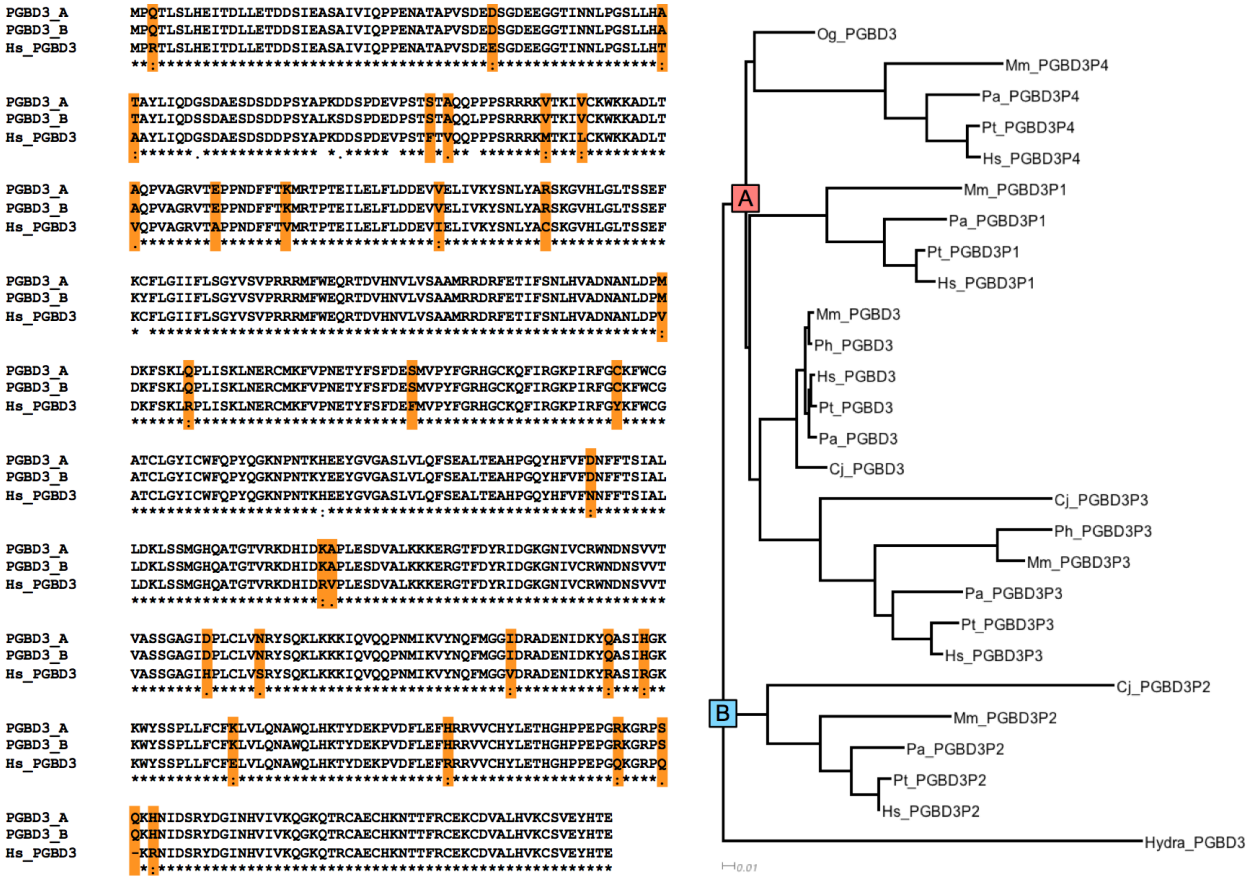



**Figure 4-8. Phylogenetic analysis of PGBD3 sequences suggests sites of mutation in domesticated human PGBD3.**

Protein sequences of all primate PGBD3, PGBD3 pseudogenes, galago PGBD3, and hydra PGBD3 were submitted to PHYLIP to calculate the phylogenetic relationship between these sequences and compute probably ancestral sequences. Sequences at position A and B in the unrooted phylogenetic tree generated by PHYLIP (right panel) were aligned to human PGBD3 (left panel). Positions where sequences A and B agree but differ from human PGBD3 are highlighted in orange. Hs, *Homo sapiens*. Pt, *Pan troglodytes*. Pa, *Pon abelii*. Ph, *Papio hamadryas*. Mm, *Macaca mulatta*. Cj, *Callithrix jacchus*. Og, *Otolemur garnettii*.

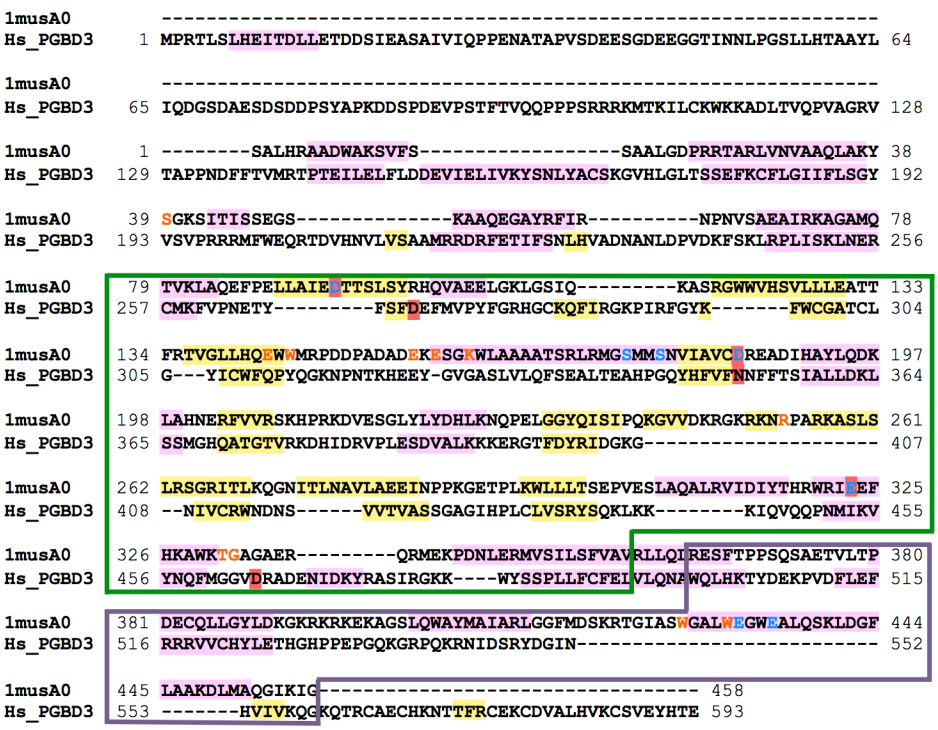



**Figure 4-9. Human PGBD3 and Tn5 have similar secondary structure.**
PSIPRED analysis shows that PGBD3 and Tn5 (PDBID 1MUS) have similar secondary structure in the RNase H core (green box) and in the N-terminal Tn5 DNA binding domains (purple box). Secondary structures for Tn5 are from crystal structure 1MUS, whereas those for PGBD3 are predicted by the PSIPRED algorithm. Magenta highlight, alpha helix. Yellow highlight, beta sheet. Red highlight, catalytic residue. Orange font, ligand coordinating residue. Blue font, Mg/Mn coordinating residue. Hs, *Homo sapiens*.

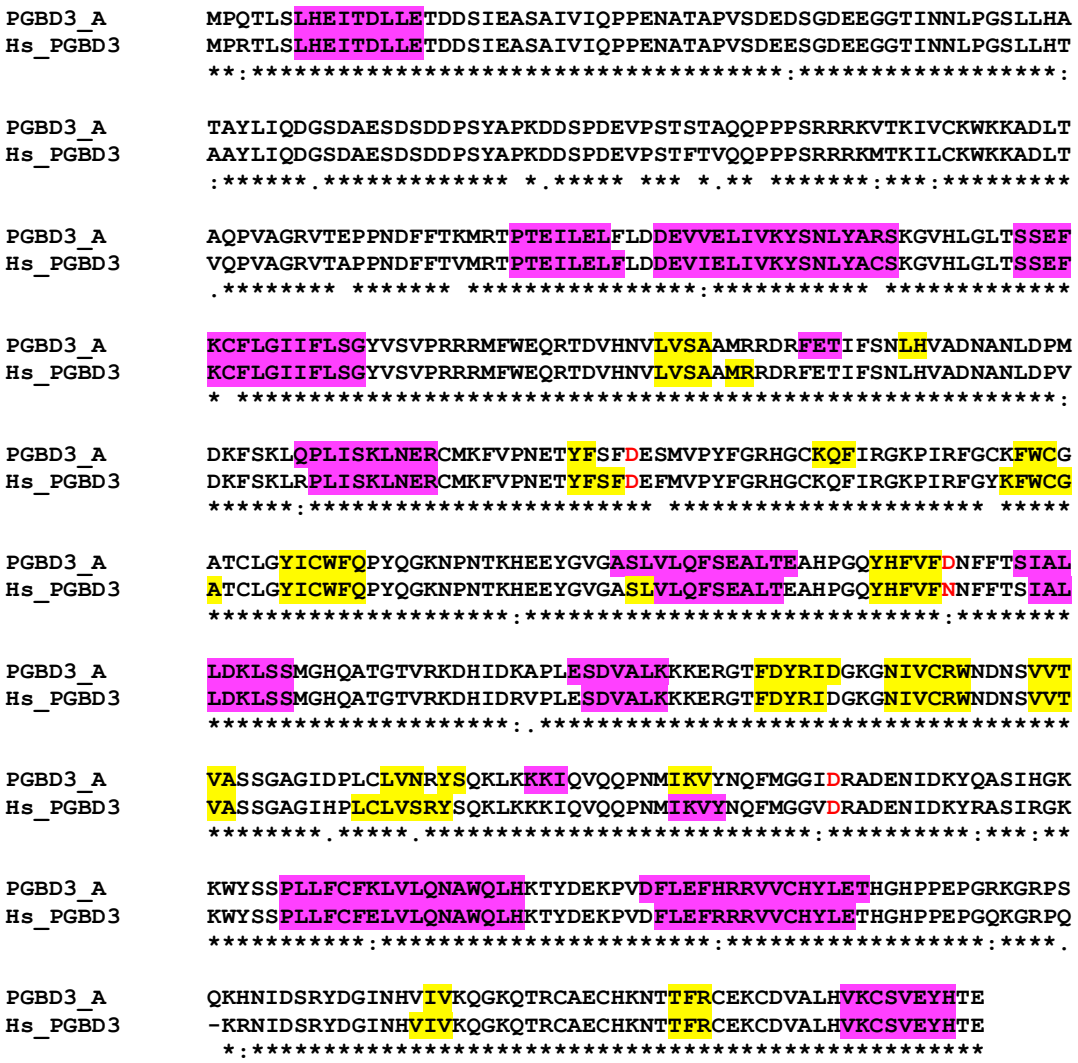



**Figure 4-10. Differences between human PGBD3 and ancestral PGBD3 sequences do not alter predicted secondary structure.**
PSIPRED predicts similar secondary structures for human PGBD3 and the deduced ancestral PGBD3_A sequence (Figure 4-8). Hs, *Homo sapiens*.

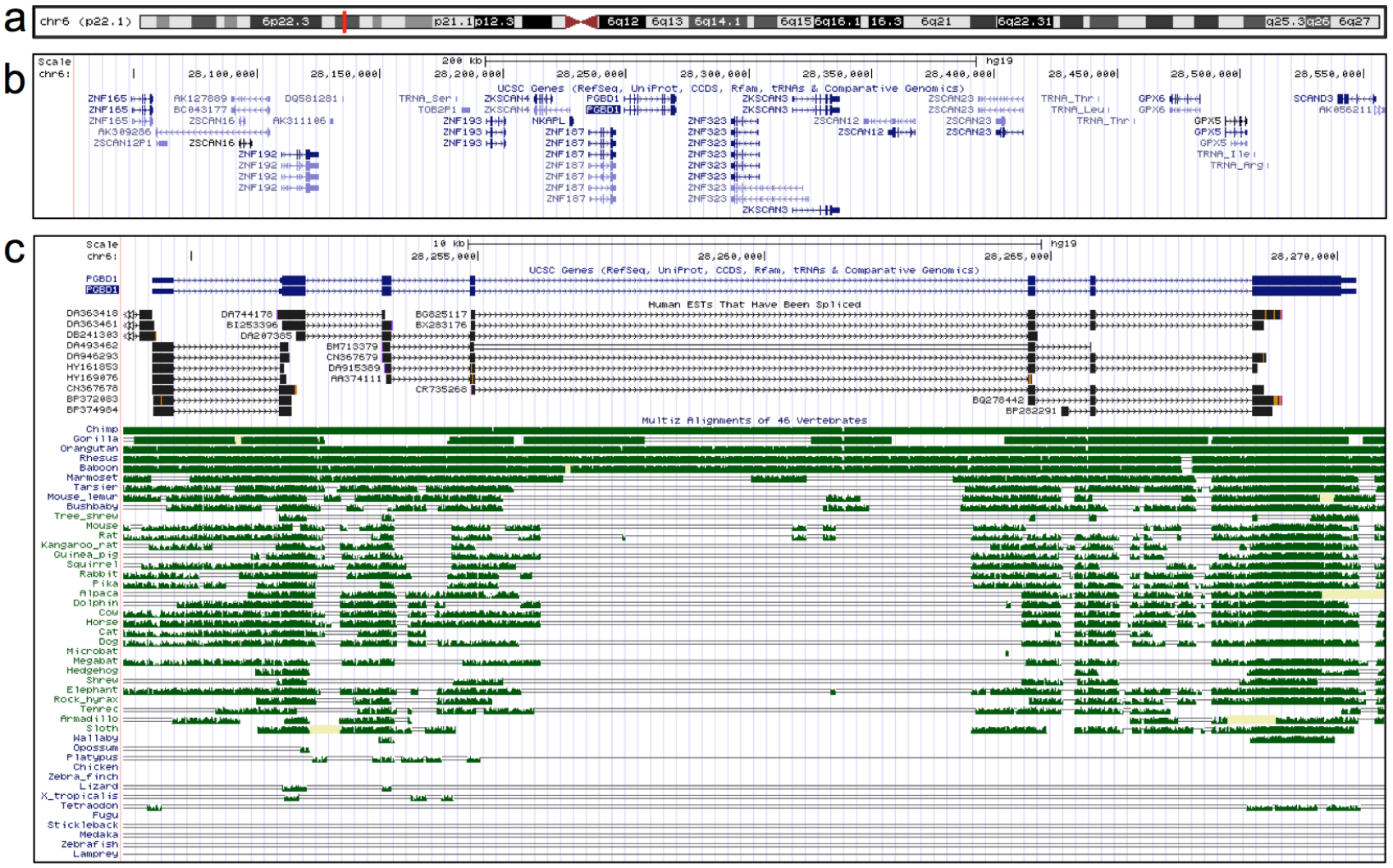

a
chr6 (p22.1)
b
Scale
chr6:
200 kb
hg19
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
c
Scale
chr6:
10 kb
hg19
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
Human ESTs That Have Been Spliced
Multiz Alignments of 46 Vertebrates

**Figure 4-11. Human PGBD1 appears to be a domesticated fusion protein surrounded by zinc finger genes.**
A. The location of PGBD1 in human chromosome 6. B. PGBD1 inserted into a region that is rich in zinc finger genes presumably resulting from repeated gene duplication events. C. Sequenced ESTs demonstrate splicing of 6 upstream exons to the 3' terminal PGBD3 transposase ORF. All 7 exons, including the piggyBac domain, are conserved throughout mammals.

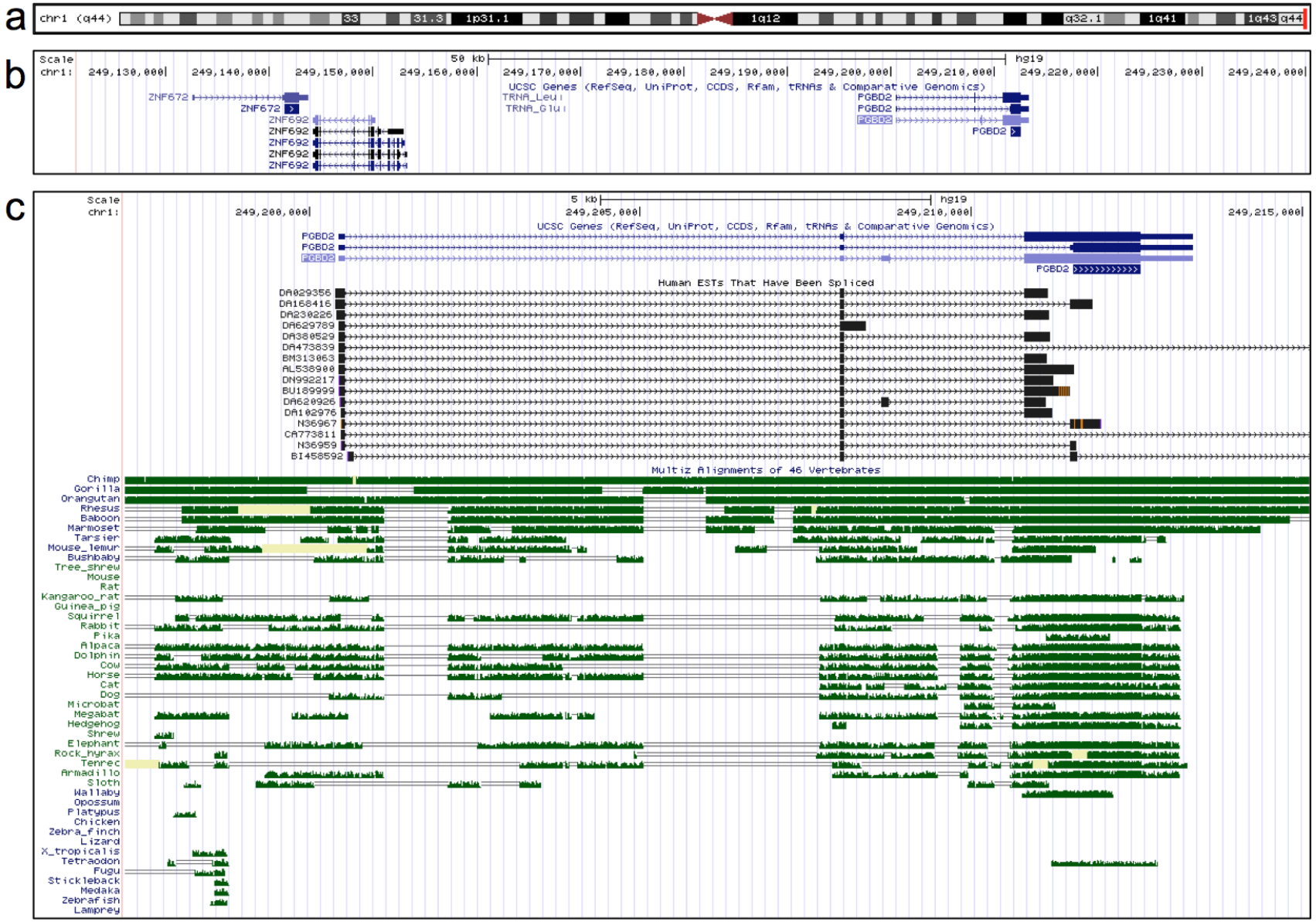

a
chr1 (q44)
33
31.3
1p31.1
1q12
q32.1
1q41
1q43
q44
b
Scale
chr1:
50 kb
hg19
249,130,000
249,140,000
249,150,000
249,160,000
249,170,000
249,180,000
249,190,000
249,200,000
249,210,000
249,220,000
249,230,000
249,240,000
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
ZNF672
ZNF692
TRNA_Leu
TRNA_Glu
PGBD2
c
Scale
chr1:
5 kb
hg19
249,200,000
249,205,000
249,210,000
249,215,000
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
PGBD2
Human ESTs That Have Been Spliced
DA029356
DA168416
DA230226
DA629789
DA380529
DA473839
BM313063
AL538900
DN992217
BU189999
DA620926
DA102976
N36967
CA773811
N36959
BI458592
Multiz Alignments of 46 Vertebrates
Chimp
Gorilla
Orangutan
Rhesus
Baboon
Marmoset
Tarsier
Mouse_lemur
Bushbaby
Tree_shrew
Mouse
Rat
Kangaroo_rat
Guinea_pig
Squirrel
Rabbit
Pika
Alpaca
Dolphin
Cow
Horse
Cat
Dog
Microbat
Megabat
Hedgehog
Shrew
Elephant
Rock_hyrax
Tenrec
Armadillo
Sloth
Wallaby
Opossum
Platypus
Chicken
Zebra_finch
Lizard
X_tropicalis
Tetraodon
Fugu
Stickleback
Medaka
Zebrafish
Lamprey

**Figure 4-12. Human PGBD2 encodes a fusion protein near Chromosome 1 telomeres.**
A. PGBD2 is located near the end of human chromosome 1p. B. PGBD2 inserted just 20 kb from the end of the chromosome 1p assembly. C. Sequenced ESTs demonstrate splicing of 2 largely noncoding upstream exons to the 3' terminal PGBD3 transposase ORF. All 3 exons, including the piggyBac domain, are conserved in most mammals.

| Accession | Gene name | Species | Max score | Total Score | Query coverage | E-value | Max identity |
|---|---|---|---|---|---|---|---|
| XP_002155162.1 | Hm_PGBD3 | *Hydra magnipapilata* | 1228 | 1228 | 98% | 0 | 100% |
| AFH27349.1 | PGBD3 | *Macaca mulatta* | 841 | 841 | 97% | 0 | 67% |
| AAH28954.2 | PGBD3 | *Homo sapiens* | 840 | 840 | 97% | 0 | 67% |
| NP_736609.2 | PGBD3 | *Homo sapiens* | 839 | 839 | 97% | 0 | 67% |
| EAW93095.1 | CSB-PGBD3 | *Homo sapiens* | 848 | 848 | 97% | 0 | 67% |
| AAH34479.1 | PGBD3 | *Homo sapiens* | 848 | 848 | 97% | 0 | 67% |
| BAE02063.1 | CSB-PGBD3 (partial) | *Macaca fascicularis* | 833 | 833 | 97% | 0 | 66% |
| AFE70772.1 | PGBD3 (partial) | *Macaca mulatta* | 754 | 754 | 85% | 0 | 69% |
| BAE90124.1 | PGBD3 (partial) | *Macaca fascicularis* | 741 | 741 | 81% | 0 | 70% |
| CAB61402.1 | PGBD3 (partial) | *Homo sapiens* | 582 | 582 | 61% | 0 | 74% |
| XP_003803632.1 | PGBD2 | *Otolemur garnettii* | 443 | 443 | 97% | 4.00E-145 | 40% |
| EHH50838.1 | PGBD2 | *Macaca fascicularis* | 439 | 439 | 93% | 2.00E-143 | 41% |
| EHH15848.1 | PGBD2 | *Macaca mulatta* | 438 | 438 | 93% | 3.00E-143 | 41% |
| NP_001245039.1 | PGBD2 | *Macaca mulatta* | 438 | 438 | 93% | 3.00E-143 | 41% |
| XP_002723800.1 | PGBD1-like | *Oryctolagus cuniculus* | 437 | 437 | 96% | 5.00E-143 | 41% |
| NP_733843.1 | PGBD2 | *Homo sapiens* | 437 | 437 | 93% | 6.00E-143 | 41% |
| XP_002809214.1 | PGBD2-like | *Pongo abelii* | 437 | 437 | 93% | 7.00E-143 | 41% |
| BAG60513.1 | PGBD2 | *Homo sapiens* | 437 | 437 | 93% | 7.00E-143 | 41% |
| XP_002809215.1 | PGBD2 | *Pongo abelii* | 437 | 437 | 93% | 8.00E-143 | 41% |
| EFB16732.1 | PGBD2 | *Ailuropoda melanoleuca* | 435 | 435 | 97% | 3.00E-142 | 41% |
| DAA27562.1 | PGBD1-like | *Bos taurus* | 435 | 435 | 97% | 5.00E-142 | 40% |
| XP_002930641.1 | PGBD2-like | *Ailuropoda melanoleuca* | 434 | 434 | 97% | 5.00E-141 | 41% |
| XP_003404993.1 | PGBD2 | *Loxodonta africana* | 437 | 437 | 97% | 7.00E-140 | 40% |
| XP_003768735.1 | PGBD2 | *Sarcophilus harrisii* | 429 | 429 | 96% | 2.00E-139 | 39% |
| XP_003639588.1 | PGBD2 | *Canis lupus familiaris* | 416 | 416 | 93% | 3.00E-134 | 39% |
| AES10976.1 | PGBD2 | *Mustela putorius furo* | 411 | 411 | 97% | 8.00E-133 | 38% |
| AFE70947.1 | PGBD2 | *Macaca mulatta* | 395 | 395 | 75% | 1.00E-128 | 43% |
| EHJ77726.1 | PGBD | *Danaus Plexippus* | 345 | 345 | 61% | 2.00E-109 | 46% |
| XP_001807508.1 | PGBD | *Tribolium castaneum* | 324 | 371 | 73% | 6.00E-101 | 53% |
| XP_002165667.1 | PGBD3 (partial) | *Hydra magnipapilata* | 295 | 295 | 37% | 2.00E-93 | 69% |
| XP_003770432.1 | PGBD1-like | *Sarcophilus harrisii* | 314 | 314 | 74% | 2.00E-93 | 39% |

**Table 4-1. The hydra PGBD3 gene is most similar to primate PGBD3.**
A protein BLAST search indicates that the hydra piggyBac element is most closely related to primate piggyBacs, especially PGBD3.

| | Hm_PGBD3 | Hs_PGBD1 | Hs_PGBD2 | Hs_PGBD3 | Hs_PGBD4 | Hs_PGBD5 | Tn_PGB |
|---|---|---|---|---|---|---|---|
| Hm_PGBD3 | | 29 | 38 | 66 | 16 | 16 | 14 |
| Hs_PGBD1 | 29 | | 50 | 35 | 18 | 10 | 15 |
| Hs_PGBD2 | 38 | 50 | | 42 | 21 | 12 | 10 |
| Hs_PGBD3 | 66 | 35 | 42 | | 16 | 16 | 16 |
| Hs_PGBD4 | 16 | 18 | 21 | 16 | | 14 | 19 |
| Hs_PGBD5 | 16 | 10 | 12 | 16 | 14 | | 15 |
| Tn_PGB | 14 | 15 | 10 | 16 | 19 | 15 | |

**Table 4-2. Identity of piggyBac transposase sequences.**
ClustalW2 was used to align human PGBD3 and PGBD3Ps with human PGBD1, PGBD2, PGBD4, PGBD5, the reassembled galago PGBD3, hydra PGBD3, and *Trichoplusia ni* piggyBac. Numbers are the percent identity of each pairwise alignment reported by ClustalW2; darker red shading indicates greater identity. Hm, *Hydra magnipapillata*. Hs, *Homo sapiens*. Tn, *Trichoplusia ni*.

| Accession | Hydra scaffold | Max score | Total score | Query coverage | E value | Max Identity |
|---|---|---|---|---|---|---|
| NW_002162101.1 | scf_1101284939372 | 5472 | 5472 | 100% | 0 | 100% |
| NW_002119966.1 | scf_1101284901420 | 3493 | 4080 | 98% | 0 | 93% |
| NW_002166009.1 | scf_1101284942889 | 3371 | 3970 | 98% | 0 | 92% |
| NW_002144211.1 | scf_1101284923263 | 2084 | 2678 | 72% | 0 | 90% |
| NW_002108756.1 | scf_1101284891328 | 1759 | 1759 | 40% | 0 | 93% |
| NW_002163931.1 | scf_1101284941019 | 1707 | 3411 | 83% | 0 | 95% |
| NW_002152556.1 | scf_1101284930780 | 1600 | 3743 | 85% | 0 | 95% |
| NW_002123922.1 | scf_1101284904982 | 1295 | 3457 | 98% | 0 | 89% |
| NW_002159204.1 | scf_1101284936765 | 1151 | 2160 | 48% | 0 | 95% |
| NW_002120137.1 | scf_1101284901575 | 1136 | 1717 | 47% | 0 | 100% |
| NW_002157288.1 | scf_1101284935040 | 1092 | 1488 | 37% | 0 | 93% |
| NW_002160687.1 | scf_1101284938100 | 791 | 2063 | 46% | 0 | 96% |
| NW_002149837.1 | scf_1101284928333 | 512 | 729 | 17% | 2.00E-140 | 93% |

**Table 4-3. The hydra genome has 13 regions with significant homology to hydra PGBD3.**
A BLAST search of the hydra genome with the hydra PGBD3 sequence as query identified 13 regions with at least partial homology to PGBD3.

| Accession | Gene Name | Max Score | Total Score | Query coverage | E value | Max Identity |
|---|---|---|---|---|---|---|
| NP_115896.1 | PGBD1 | 596 | 596 | 100% | 0 | 100% |
| NP_443155.1 | **SCAND3** | 158 | 158 | 60% | 4.00E-44 | 50% |
| NP_003430.1 | ZKSCAN1 | 154 | 154 | 96% | 9.00E-44 | 41% |
| NP_006289.2 | **ZNF192** | 150 | 150 | 98% | 1.00E-42 | 38% |
| NP_079507.1 | ZSCAN16 | 145 | 145 | 69% | 1.00E-41 | 45% |
| NP_001128688.1 | **ZNF323 isoform 1** | 141 | 141 | 53% | 5.00E-40 | 53% |
| NP_115723.1 | ZNF397 isoform 2 | 137 | 137 | 53% | 2.00E-39 | 48% |
| NP_001156863.1 | **ZSCAN12** | 142 | 142 | 48% | 2.00E-39 | 55% |
| NP_079445.1 | ZNF167 isoform 2 | 135 | 135 | 86% | 9.00E-39 | 39% |
| NP_061121.2 | ZNF167 isoform 1 | 141 | 141 | 96% | 1.00E-38 | 38% |
| NP_001012458.1 | **ZSCAN23** | 135 | 135 | 45% | 5.00E-38 | 58% |
| NP_001128650.1 | ZNF397 isoform 1 | 134 | 134 | 46% | 4.00E-37 | 55% |
| NP_001020026.1 | ZNF197 isoform 2 | 130 | 130 | 84% | 7.00E-37 | 38% |
| NP_003438.1 | **ZNF165** | 133 | 133 | 46% | 7.00E-37 | 54% |
| NP_055384.1 | ZKSCAN5 | 135 | 135 | 93% | 2.00E-36 | 40% |
| NP_005732.2 | ZNF263 | 132 | 132 | 89% | 6.00E-36 | 35% |
| NP_008896.2 | ZNF24 | 129 | 129 | 44% | 7.00E-36 | 55% |
| NP_004211.1 | ZNF213 | 126 | 126 | 39% | 3.00E-34 | 56% |
| NP_008922.1 | ZNF197 isoform 1 | 129 | 129 | 92% | 3.00E-34 | 36% |
| NP_001027463.1 | ZNF174 isoform b | 119 | 119 | 35% | 3.00E-33 | 57% |
| NP_061983.2 | **ZKSCAN4** | 122 | 122 | 91% | 1.00E-32 | 37% |
| NP_055334.2 | ZNF232 | 121 | 121 | 42% | 1.00E-32 | 57% |
| NP_003441.1 | ZNF174 isoform a | 119 | 119 | 35% | 5.00E-32 | 57% |
| NP_077819.2 | **ZKSCAN3 isoform 1** | 118 | 118 | 77% | 3.00E-31 | 41% |
| NP_115540.2 | ZNF394 | 116 | 152 | 73% | 1.00E-30 | 52% |
| NP_001106205.1 | ZSCAN30 | 115 | 115 | 47% | 2.00E-30 | 51% |
| NP_001007170.1 | ZNF483 | 112 | 112 | 30% | 3.00E-30 | 60% |
| NP_003446.2 | ZNF202 | 115 | 115 | 86% | 6.00E-30 | 35% |
| NP_666019.1 | ZSCAN21 | 113 | 113 | 32% | 8.00E-30 | 62% |
| NP_001253962.1 | MZF1 | 111 | 111 | 47% | 8.00E-30 | 48% |
| NP_001018854.2 | **ZNF187** | 113 | 113 | 43% | 9.00E-30 | 53% |

**Table 4-4. The N-terminus of PGBD1 is homologous to zinc finger proteins.**
The 6 exons upstream of the 3' terminal PGBD1 exon are predicted to encode the 289 N-terminal residues of the conserved PGBD1 fusion protein. A BLAST search with these 289 residues identified zinc finger proteins, including many located in the 200 kb window shown in Figure 4-10 (bold gene names).

# Chapter 5: Prospectus

When I started work on the CSB-PGBD3 fusion protein, little was known about its function. CSB-PGBD3 is startlingly well conserved and continues to be expressed in all anthropoid primates. Our genome-wide analysis of CSB-PGBD3 DNA binding patterns and gene regulation in CSB-null cells has begun to reveal where and how CSB-PGBD3 acts in transcription. CSB-PGBD3 interacts with transcription factors that long predate the insertion of PGBD3 into primate genomes. When CSB-PGBD3 is expressed in the absence of CSB, the fusion protein causes up-regulation of a set of genes related to the late phase of the innate immune response that is normally regulated by unphosphorylated STATs. Coexpression of full-length CSB with CSB-PGBD3, as in normal cells, represses the innate immune response and instead up- and down-regulates a substantially different set of genes.This leads to a new hypothesis that induction of an innate immune response by CSB-PGBD3, or a lack of appropriate down-regulation of such a response resulting from loss of CSB function, could play a role in the pathology of Cockayne syndrome.

CSB and CSB-PGBD3 share the same N-terminal domain, and this could account for down-regulation of the innate immune response when both proteins are expressed. CSB-PGBD3 interacts with the AP-1 transcription factor c-Jun and with RNAPII through the CSB N-terminal domain (Chapter 3). Thus, CSB might compete with CSB-PGBD3 for AP-1, RNAPII, and other binding partners.

To test these and other questions raised by my doctoral research, I propose several additional experiments and lines of investigation:

**1. Inflammation in Cockayne syndrome.** We have shown that inflammatory or innate immune pathways are perturbed both in CS1AN cells [6] which continue to express CSB-PGBD3 [14] and in UVSS1KO cells stably transfected with CSB-PGBD3 (Chapter 2). To see if these results extend to CS patients, we need to test patient cells or blood for markers of inflammation such as interferons and other pro-inflammatory cytokines. If an abnormal innate immune or inflammatory response is detected in patients, targeted anti-inflammatory drugs may be used to treat some of the symptoms of the disease.

**2. Sequestration of CSB in Cockayne syndrome.** Are CS symptoms caused by unrepaired DNA damage or by a lack of CSB required for other functions such as transcription initiation or elongation, gene regulation, or signaling? Anindya et al. [8] have shown that CSB lacking its ubiquitin binding domain becomes trapped in unproductive TC-NER complexes. It remains to be determined whether mutations in other TC-NER proteins can also lead to assembly of unproductive TC-NER complexes.

If so, this could explain why mutations in CSA and rare variants of XPB, D, F, and G cause not just DNA repair defects, as in xeroderma pigmentosum, but Cockayne syndrome as well. Photobleaching experiments with fluorescently tagged CSB (as in [8]) and chromatin partitioning experiments [89] for untagged CSB could be performed in cells expressing mutant forms of CSA, XPB, XPD, XPF, or XPG to see if CSB is being trapped in repair complexes.

**3. Mechanisms of CSB and CSB-PGBD3 in gene regulation.** The mechanism by which CSB and CSB-PGBD3 affect gene expression is not clear. CSB is known to enhance transcriptional elongation by both RNAPI [3,142] and RNAPII [4,165], and expression array analysis of CS1AN cells suggests that CSB also has a role in chromatin remodeling [6]. Tandem affinity purification of CSB complexes would reveal whether CSB interacts directly with chromatin remodeling factors, transcription factors, or other proteins related to transcription. The same could be done with the CSB-PGBD3 fusion protein to determine how similar CSB and CSB-PGBD3 complexes are, and whether CSB-PGBD3 may impersonate CSB in CS cells. Finally, co-expression CSB and CSB-PGBD3, where only one is affinity tagged, could show whether CSB affects the proteins associated with CSB-PGBD3, or vice versa, and if CSB and CSB-PGBD3 associate with each other.

**4. Mechanism of CSB-PGBD3 in DNA repair.** Our host cell reactivation (HCR) experiments showed that CSB-PGBD3 expression enhanced repair of UV damage in CSB-null cells, and that the enhancement of repair is additive when CSB-PGBD3 and CSB are coexpressed (Chapter 2). This indicates that CSB-PGBD3 and full-length CSB could have either synergistic or non-redundant roles in repair of some types of DNA damage. Chromatin fractionation experiments [89] in UVSS1KO cells expressing CSB-PGBD3 could be used to test for binding of CSB-PGBD3 to damaged DNA after exposure to UV, and IPs could be used to see whether UV damage induces specific interactions between CSB-PGBD3 and TCR proteins.

If so, CSB-PGBD3 might facilitate UV repair simply by supplying more of the CSB N-terminus to cells in need of this domain. And in this case, expression of chimeric CSB-eGFP or CSB-LacI fusion proteins in which CSB exons 1-5 are fused to different C-terminal domains, could augment UV damage repair as well. This could be easily tested in CSB-null UVSS cells expressing these chimeric proteins (used in Chapter 3) with and without CSB.

**5. The role of CSB-PGBD3 sequence and binding in alternative splicing.** Unlike other DNA transposons which act as insertional mutagens, piggyBac elements (including PGBD3) typically avoid this handicap by functioning as alternative 3' exons

that allow continued expression of the host gene. Thus alternative splicing and efficient translation enable the CSB locus to maintain balanced expression of full-length CSB and the CSB-PGBD3 fusion protein. This could explain why the newborn CSB-PGBD3 fusion protein was initially tolerated after PGBD3 insertion. ChIP-seq analysis of CSB-PGBD3 showed that the fusion protein and presumably the solitary PGBD3 transposase bind strongly to the terminal inverted repeats and palindromic binding sites of the PGBD3 element flanking the PGBD3 ORF. CSB-PGBD3 binding to these sites could potentially regulate alternative splicing of the PGBD3 ORF by interacting with splicing factors, or by retarding or even obstructing RNAPII elongation [125]. This could be tested by cloning the PGBD3 5' region — including the splice acceptor site, TIR, and two PGBD3 palindromes — into a simplified alternative splicing reporter construct.

**6. CSB-PGBD3 expression in a non-anthropoid system.** The PGBD3 insertion in CSB is only found in anthropoid primates, and neither CSB-PGBD3 fusion protein nor MER85 elements are found outside of this clade. In view of our discovery that the CSB-PGBD3 fusion protein affects gene expression near sites of interaction with conserved transcription factors, but not when bound directly to MER85 elements, we may be able to use non-simian primate or even murine lines as MER85-null host cells to determine whether similar transcription changes can be induced in cells without MER85 binding sites. The human and mouse N-terminal CSB domains are very similar but not identical. Thus if expression of hCSB-PGBD3, mCSB-PGBD3, or a chimeric mCSB-eGFP or mCSB-LacI fusion protein can induce a similar innate immune response in cells that lack MER85 elements, we would have even stronger evidence that CSB-PGBD3 regulates gene expression primarily through interactions with other transcription factors and/or CSB-related complexes, and not through MER85-bound complexes.

**7. Characterization of the hydra PGBD3.** The PGBD3 ORF in *Hydra magnipapillata* appears to have been recently active in hydra, and unlike the primate PGBD3 sequences, still retains all three of its catalytic aspartate residues (Chapter 4). Thus, it's possible that the hydra PGBD3 is still capable of at least some steps in the overall transposition reaction. If so, chimeric constructs with parts of the human and hydra PGBD3 could be used to determine what domains of PGBD3 have been altered during domestication of the primate version.

**8. What selective advantage did the original PGBD3 insertion in CSB intron 5 confer on our anthropoid ancestor?** Viral infection is a major evolutionary driver of adaptations that alert cells to infection through innate immune pathways [166,167]. The ability of the CSB-PGBD3 fusion protein to induce an innate immune response (Chapter 2) suggests that CSB-PGBD3, and possibly the solitary PGBD3 transposase, could help to protect cells from pathogens. To test this hypothesis, cell survival could be assayed

after infection with a panel of viruses. Ideally, these assays would be performed in cells with or without both CSB-PGBD3 and the solitary PGBD3 transposase, effectively recreating the genotype of cells before and after the conserved PGBD3 insertion. This could be achieved with shRNA or siRNA knockdown of the PGBD3 domain shared by CSB-PGBD3 and the solitary PGBD3 transposase. Alternately, the PGBD3 insertion could be removed using an adeno-associated virus gene targeting [168]. However, knockdowns with siRNA have already been used to reduce expression of CSB-PGBD3 without affecting expression of CSB [37], and knockdowns would have the added benefit of quick application to multiple cell types.

Knockdown of PGBD3 could also be used to test changes to gene expression and DNA repair, and complementation with siRNA-resistant CSB-PGBD3 or solitary PGBD3 transposase could be used to deconvolute the contributions of each protein to repair, transcription, and viral resistance.

# References


1. Fousteri M, Vermeulen W, van Zeeland AA, Mullenders LHF (2006) Cockayne syndrome A and B proteins differentially regulate recruitment of chromatin remodeling and repair factors to stalled RNA polymerase II in vivo. Mol Cell 23: 471–482. doi:10.1016/j.molcel.2006.06.029.

2. Laugel V, Dalloz C, Durand M, Sauvanaud F, Kristensen U, et al. (2010) Mutation update for the CSB/ERCC6 and CSA/ERCC8 genes involved in Cockayne syndrome. Hum Mutat 31: 113–126. doi:10.1002/humu.21154.

3. Yuan X, Feng W, Imhof A, Grummt I, Zhou Y (2007) Activation of RNA polymerase I transcription by cockayne syndrome group B protein and histone methyltransferase G9a. Mol Cell 27: 585–595. doi:10.1016/j.molcel.2007.06.021.

4. Tantin D, Kansal A, Carey M (1997) Recruitment of the putative transcription-repair coupling factor CSB/ERCC6 to RNA polymerase II elongation complexes. Mol Cell Biol 17: 6803–6814.

5. Filippi S, Latini P, Frontini M, Palitti F, Egly J-M, et al. (2008) CSB protein is (a direct target of HIF-1 and) a critical mediator of the hypoxic response. EMBO J 27: 2545–2556. doi:10.1038/emboj.2008.180.

6. Newman JC, Bailey AD, Weiner AM (2006) Cockayne syndrome group B protein (CSB) plays a general role in chromatin maintenance and remodeling. Proc Natl Acad Sci USA 103: 9613–9618. doi:10.1073/pnas.0510909103.

7. Groisman R, Kuraoka I, Chevallier O, Gaye N, Magnaldo T, et al. (2006) CSA-dependent degradation of CSB by the ubiquitin-proteasome pathway establishes a link between complementation factors of the Cockayne syndrome. Genes Dev 20: 1429–1434. doi:10.1101/gad.378206.

8. Anindya R, Mari P-O, Kristensen U, Kool H, Giglia-Mari G, et al. (2010) A ubiquitin-binding domain in Cockayne syndrome B required for transcription-coupled nucleotide excision repair. Mol Cell 38: 637–648. doi:10.1016/j.molcel.2010.04.017.

9. Gray LT, Weiner AM (2010) Ubiquitin recognition by the Cockayne syndrome group B protein: binding will set you free. Mol Cell 38: 621–622. doi:10.1016/j.molcel.2010.05.025.

10. Ito S, Kuraoka I, Chymkowitch P, Compe E, Takedachi A, et al. (2007) XPG stabilizes TFIIH, allowing transactivation of nuclear receptors: implications for Cockayne syndrome in XP-G/CS patients. Mol Cell 26: 231–243. doi:10.1016/j.molcel.2007.03.013.

11. Nakazawa Y, Sasaki K, Mitsutake N, Matsuse M, Shimada M, et al. (2012) Mutations in UVSSA cause UV-sensitive syndrome and impair RNA polymerase IIo processing in transcription-coupled nucleotide-excision repair. Nat Genet 44: 586–592. doi:10.1038/ng.2229.

12. Schwertman P, Lagarou A, Dekkers DHW, Raams A, van der Hoek AC, et al. (2012) UV-sensitive syndrome protein UVSSA recruits USP7 to regulate transcription-coupled repair. Nat Genet 44: 598–602. doi:10.1038/ng.2230.

13. Zhang X, Horibata K, Saijo M, Ishigami C, Ukai A, et al. (2012) Mutations in UVSSA cause UV-sensitive syndrome and destabilize ERCC6 in transcription-coupled DNA repair. Nat Genet 44: 593–597. doi:10.1038/ng.2228.

14. Newman JC, Bailey AD, Fan H-Y, Pavelitz T, Weiner AM (2008) An Abundant Evolutionarily Conserved CSB-PiggyBac Fusion Protein Expressed in Cockayne Syndrome. PLoS Genet 4: e1000031. doi:10.1371/journal.pgen.1000031.

15. Horibata K, Iwamoto Y, Kuraoka I, Jaspers NGJ, Kurimasa A, et al. (2004) Complete absence of Cockayne syndrome group B gene product gives rise to UV-sensitive syndrome but not Cockayne syndrome. Proc Natl Acad Sci USA 101: 15410–15415. doi:10.1073/pnas.0404587101.

16. van der Horst GT, van Steeg H, Berg RJ, van Gool AJ, de Wit J, et al. (1997) Defective transcription-coupled repair in Cockayne syndrome B mice is associated with skin cancer predisposition. Cell 89: 425–435.

17. Murai M, Enokido Y, Inamura N, Yoshino M, Nakatsu Y, et al. (2001) Early postnatal ataxia and abnormal cerebellar development in mice lacking Xeroderma pigmentosum Group A and Cockayne syndrome Group B DNA repair genes. Proc Natl Acad Sci USA 98: 13379–13384. doi:10.1073/pnas.231329598.

18. Laposa RR, Huang EJ, Cleaver JE (2007) Increased apoptosis, p53 up-regulation, and cerebellar neuronal degeneration in repair-deficient Cockayne syndrome mice. Proc Natl Acad Sci USA 104: 1389–1394. doi:10.1073/pnas.0610619104.

19. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921. doi:10.1038/35057062.

20. Feschotte C, Pritham EJ (2007) DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet 41: 331–368. doi:10.1146/annurev.genet.40.110405.090448.

21. Wang T, Zeng J, Lowe CB, Sellers RG, Salama SR, et al. (2007) Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. Proc Natl Acad Sci USA 104: 18613–

18618. doi:10.1073/pnas.0703637104.

22. Lynch VJ, Leclerc RD, May G, Wagner GP (2011) Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. Nat Genet 43: 1154–1159. doi:10.1038/ng.917.

23. Baudry C, Malinsky S, Restituito M, Kapusta A, Rosa S, et al. (2009) PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements in the ciliate Paramecium tetraurelia. Genes Dev 23: 2478–2483. doi:10.1101/gad.547309.

24. McClintock B (1951) Chromosome organization and genic expression. Cold Spring Harb Symp Quant Biol 16: 13–47.

25. Britten RJ, Davidson EH (1969) Gene regulation for higher cells: a theory. Science 165: 349–357.

26. Bailey AD, Gray LT, Pavelitz T, Newman JC, Horibata K, et al. (2012) The conserved Cockayne syndrome B-piggyBac fusion protein (CSB-PGBD3) affects DNA repair and induces both interferon-like and innate antiviral responses in CSB-null cells. DNA Repair (Amst). doi:10.1016/j.dnarep.2012.02.004.

27. Beerens N, Hoeijmakers JHJ, Kanaar R, Vermeulen W, Wyman C (2005) The CSB protein actively wraps DNA. J Biol Chem 280: 4722–4729. doi:10.1074/jbc.M409147200.

28. Citterio E, Van Den Boom V, Schnitzler G, Kanaar R, Bonte E, et al. (2000) ATP-dependent chromatin remodeling by the Cockayne syndrome B DNA repair-transcription-coupling factor. Mol Cell Biol 20: 7643–7653.

29. Nardo T, Oneda R, Spivak G, Vaz B, Mortier L, et al. (2009) A UV-sensitive syndrome patient with a specific CSA mutation reveals separable roles for CSA in response to UV and oxidative DNA damage. Proc Natl Acad Sci USA 106: 6209–6214. doi:10.1073/pnas.0902113106.

30. Harreman M, Taschner M, Sigurdsson S, Anindya R, Reid J, et al. (2009) Distinct ubiquitin ligases act sequentially for RNA polymerase II polyubiquitylation. Proc Natl Acad Sci USA 106: 20705–20710. doi:10.1073/pnas.0907052106.

31. Fousteri M, Mullenders LHF (2008) Transcription-coupled nucleotide excision repair in mammalian cells: molecular mechanisms and biological effects. Cell Res 18: 73–84. doi:10.1038/cr.2008.6.

32. Hashimoto S, Suga T, Kudo E, Ihn H, Uchino M, et al. (2008) Adult-onset neurological degeneration in a patient with Cockayne syndrome and a null mutation in the CSB gene. J Invest Dermatol 128: 1597–1599. doi:10.1038/sj.jid.5701210.

33. Laugel V, Dalloz C, Stary A, Cormier-Daire V, Desguerre I, et al. (2008) Deletion of 5' sequences of the CSB gene provides insight into the pathophysiology of Cockayne syndrome. Eur J Hum Genet 16: 320–327. doi:10.1038/sj.ejhg.5201991.

34. Nance MA, Berry SA (1992) Cockayne syndrome: review of 140 cases. Am J Med Genet 42: 68–84. doi:10.1002/ajmg.1320420115.

35. Colella S, Nardo T, Botta E, Lehmann AR, Stefanini M (2000) Identical mutations in the CSB gene associated with either Cockayne syndrome or the DeSanctis-cacchione variant of xeroderma pigmentosum. Hum Mol Genet 9: 1171–1175.

36. Spivak G, Hanawalt PC (2006) Host cell reactivation of plasmids containing oxidative DNA lesions is defective in Cockayne syndrome but normal in UV-sensitive syndrome fibroblasts. DNA Repair (Amst) 5: 13–22. doi:10.1016/j.dnarep.2005.06.017.

37. Horibata K, Saijo M, Bay MN, Lan L, Kuraoka I, et al. (2011) Mutant Cockayne syndrome group B protein inhibits repair of DNA topoisomerase I-DNA covalent complex. Genes Cells 16: 101–114. doi:10.1111/j.1365-2443.2010.01467.x.

38. Itoh T, Yamaizumi M (2000) UVs syndrome: establishment and characterization of fibroblastic cell lines transformed with simian virus 40 DNA. J Invest Dermatol 114: 101–106. doi:10.1046/j.1523-1747.2000.00843.x.

39. Newman JC, Weiner AM (2005) L2L: a simple tool for discovering the hidden significance in microarray expression data. Genome Biol 6: R81. doi:10.1186/gb-2005-6-9-r81.

40. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, et al. (2010) GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol 28: 495–501. doi:10.1038/nbt.1630.

41. Cheon H, Stark GR (2009) Unphosphorylated STAT1 prolongs the expression of interferon-induced immune regulatory genes. Proc Natl Acad Sci USA 106: 9373–9378. doi:10.1073/pnas.0903487106.

42. Malathi K, Dong B, Gale M, Silverman RH (2007) Small self-RNA generated by RNase L amplifies antiviral innate immunity. Nature 448: 816–819. doi:10.1038/nature06042.

43. Chatterjee-Kishore M, Wright KL, Ting JP, Stark GR (2000) How Stat1 mediates constitutive gene expression: a complex of unphosphorylated Stat1 and IRF1 supports transcription of the LMP2 gene. EMBO J 19: 4111–4122. doi:10.1093/emboj/19.15.4111.

44. Cheon H, Yang J, Stark GR (2011) The functions of signal transducers and

activators of transcriptions 1 and 3 as cytokine-inducible proteins. J Interferon Cytokine Res 31: 33–40. doi:10.1089/jir.2010.0100.

45. Hou Y-C, Liou K-T, Chern C-M, Wang Y-H, Liao J-F, et al. (2010) Preventive effect of silymarin in cerebral ischemia-reperfusion-induced brain injury in rats possibly through impairing NF-κB and STAT-1 activation. Phytomedicine 17: 963–973. doi:10.1016/j.phymed.2010.03.012.

46. Stephanou A, Scarabelli TM, Brar BK, Nakanishi Y, Matsumura M, et al. (2001) Induction of apoptosis and Fas receptor/Fas ligand expression by ischemia/reperfusion in cardiac myocytes requires serine 727 of the STAT-1 transcription factor but not tyrosine 701. J Biol Chem 276: 28340–28347. doi:10.1074/jbc.M101177200.

47. Wilkins C, Gale M (2010) Recognition of viruses by cytoplasmic sensors. Curr Opin Immunol 22: 41–47. doi:10.1016/j.coi.2009.12.003.

48. Gerlier D, Avice T (1984) Use of an automatic cell harvester in a cellular radioimmunoassay. J Immunol Methods 75: 159–166.

49. Jiang L-J, Zhang N-N, Ding F, Li X-Y, Chen L, et al. (2011) RA-inducible gene-I induction augments STAT1 activation to inhibit leukemia cell proliferation. Proc Natl Acad Sci USA 108: 1897–1902. doi:10.1073/pnas.1019059108.

50. Sauter D, Specht A, Kirchhoff F (2010) Tetherin: holding on and letting go. Cell 141: 392–398. doi:10.1016/j.cell.2010.04.022.

51. Evans DT, Serra-Moreno R, Singh RK, Guatelli JC (2010) BST-2/tetherin: a new component of the innate immune response to enveloped viruses. Trends Microbiol 18: 388–396. doi:10.1016/j.tim.2010.06.010.

52. van Boxel-Dezaire AHH, Stark GR (2007) Cell type-specific signaling in response to interferon-gamma. Curr Top Microbiol Immunol 316: 119–154.

53. van der Pluijm I, Garinis GA, Brandt RMC, Gorgels TGMF, Wijnhoven SW, et al. (2007) Impaired genome maintenance suppresses the growth hormone--insulin-like growth factor 1 axis in mice with Cockayne syndrome. PLoS Biol 5: e2. doi:10.1371/journal.pbio.0050002.

54. Niedernhofer LJ, Garinis GA, Raams A, Lalai AS, Robinson AR, et al. (2006) A new progeroid syndrome reveals that genotoxic stress suppresses the somatotroph axis. Nature 444: 1038–1043. doi:10.1038/nature05456.

55. Sawyer SL, Emerman M, Malik HS (2007) Discordant evolution of the adjacent antiretroviral genes TRIM22 and TRIM5 in mammals. PLoS Pathog 3: e197. doi:10.1371/journal.ppat.0030197.

56. Dong B, Zhou Q, Zhao J, Zhou A, Harty RN, et al. (2004) Phospholipid

scramblase 1 potentiates the antiviral activity of interferon. J Virol 78: 8983–8993. doi:10.1128/JVI.78.17.8983-8993.2004.

57. Rivieccio MA, Suh H-S, Zhao Y, Zhao M-L, Chin KC, et al. (2006) TLR3 ligation activates an antiviral response in human fetal astrocytes: a role for viperin/cig5. J Immunol 177: 4735–4741.

58. Feschotte C (2008) Transposable elements and the evolution of regulatory networks. Nat Rev Genet 9: 397–405. doi:10.1038/nrg2337.

59. Ares M, Chung JS, Giglio L, Weiner AM (1987) Distinct factors with Sp1 and NF-A specificities bind to adjacent functional elements of the human U2 snRNA gene enhancer. Genes Dev 1: 808–817.

60. Garinis GA, Uittenboogaard LM, Stachelscheid H, Fousteri M, van Ijcken W, et al. (2009) Persistent transcription-blocking DNA lesions trigger somatic growth attenuation associated with longevity. Nat Cell Biol 11: 604–615. doi:10.1038/ncb1866.

61. Garinis GA, van der Horst GTJ, Vijg J, Hoeijmakers JHJ (2008) DNA damage and ageing: new-age ideas for an age-old problem. Nat Cell Biol 10: 1241–1247. doi:10.1038/ncb1108-1241.

62. Shuai K, Liu B (2003) Regulation of JAK-STAT signalling in the immune system. Nat Rev Immunol 3: 900–911. doi:10.1038/nri1226.

63. Gough DJ, Levy DE, Johnstone RW, Clarke CJ (2008) IFNgamma signaling-does it mean JAK-STAT? Cytokine Growth Factor Rev 19: 383–394. doi:10.1016/j.cytogfr.2008.08.004.

64. Chen L-L, Yang L, Carmichael GG (2010) Molecular basis for an attenuated cytoplasmic dsRNA response in human embryonic stem cells. Cell Cycle 9: 3552–3564.

65. Brooks PJ, Cheng T-F, Cooper L (2008) Do all of the neurologic diseases in patients with DNA repair gene mutations result from the accumulation of DNA damage? DNA Repair (Amst) 7: 834–848. doi:10.1016/j.dnarep.2008.01.017.

66. Gall A, Treuting P, Elkon KB, Loo Y-M, Gale M, et al. (2012) Autoimmunity Initiates in Nonhematopoietic Cells and Progresses via Lymphocytes in an Interferon-Dependent Autoimmune Disease. Immunity 36: 120–131. doi:10.1016/j.immuni.2011.11.018.

67. Alic N, Hoddinott MP, Vinti G, Partridge L (2011) Lifespan extension by increased expression of the Drosophila homologue of the IGFBP7 tumour suppressor. Aging Cell 10: 137–147. doi:10.1111/j.1474-9726.2010.00653.x.

68. Sykora P, Wilson DM, Bohr VA (2012) Repair of persistent strand breaks in the

mitochondrial genome. Mech Ageing Dev 133: 169–175. doi:10.1016/j.mad.2011.11.003.

69. Gough SM, Slape CI, Aplan PD (2011) NUP98 gene fusions and hematopoietic malignancies: common themes and new biologic insights. Blood 118: 6247–6257. doi:10.1182/blood-2011-07-328880.

70. Ghannam G, Takeda A, Camarata T, Moore MA, Viale A, et al. (2004) The oncogene Nup98-HOXA9 induces gene transcription in myeloid cells. J Biol Chem 279: 866–875. doi:10.1074/jbc.M307280200.

71. Takeda A, Sarma NJ, Abdul-Nabi AM, Yaseen NR (2010) Inhibition of CRM1-mediated nuclear export of transcription factors by leukemogenic NUP98 fusion proteins. J Biol Chem 285: 16248–16257. doi:10.1074/jbc.M109.048785.

72. Cordaux R, Udit S, Batzer MA, Feschotte C (2006) Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. Proc Natl Acad Sci USA 103: 8101–8106. doi:10.1073/pnas.0601161103.

73. Fnu S, Williamson EA, De Haro LP, Brenneman M, Wray J, et al. (2011) Methylation of histone H3 lysine 36 enhances DNA repair by nonhomologous end-joining. Proc Natl Acad Sci USA 108: 540–545. doi:10.1073/pnas.1013571108.

74. Beck BD, Lee S-S, Williamson E, Hromas RA, Lee S-H (2011) Biochemical characterization of metnase's endonuclease activity and its role in NHEJ repair. Biochemistry 50: 4360–4370. doi:10.1021/bi200333k.

75. Wen Z, Zhong Z, Darnell JE (1995) Maximal activation of transcription by Stat1 and Stat3 requires both tyrosine and serine phosphorylation. Cell 82: 241–250.

76. Mazière C, Dantin F, Dubois F, Santus R, Mazière J (2000) Biphasic effect of UVA radiation on STAT1 activity and tyrosine phosphorylation in cultured human keratinocytes. Free Radic Biol Med 28: 1430–1437.

77. Kovarik P, Mangold M, Ramsauer K, Heidari H, Steinborn R, et al. (2001) Specificity of signaling by STAT1 depends on SH2 and C-terminal domains that regulate Ser727 phosphorylation, differentially affecting specific target gene expression. EMBO J 20: 91–100. doi:10.1093/emboj/20.1.91.

78. Yu A, Fan HY, Liao D, Bailey AD, Weiner AM (2000) Activation of p53 or loss of the Cockayne syndrome group B repair protein causes metaphase fragility of human U1, U2, and 5S genes. Mol Cell 5: 801–810.

79. Vermeulen W, Rademakers S, Jaspers NG, Appeldoorn E, Raams A, et al. (2001) A temperature-sensitive disorder in basal transcription and DNA repair in humans. Nat Genet 27: 299–303. doi:10.1038/85864.

80. Pellegrini S, John J, Shearer M, Kerr IM, Stark GR (1989) Use of a selectable marker regulated by alpha interferon to obtain mutations in the signaling pathway. Mol Cell Biol 9: 4605–4612.

81. Lim CP, Cao X (2006) Structure, function, and regulation of STAT proteins. Mol Biosyst 2: 536–550. doi:10.1039/b606246f.

82. Dauer DJ, Ferraro B, Song L, Yu B, Mora L, et al. (2005) Stat3 regulates genes common to both wound healing and cancer. Oncogene 24: 3397–3408. doi:10.1038/sj.onc.1208469.

83. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 102: 15545–15550. doi:10.1073/pnas.0506580102.

84. van den Boom V, Citterio E, Hoogstraten D, Zotter A, Egly J-M, et al. (2004) DNA damage stabilizes interaction of CSB with the transcription elongation machinery. J Cell Biol 166: 27–36. doi:10.1083/jcb.200401056.

85. van Gool AJ, Citterio E, Rademakers S, van Os R, Vermeulen W, et al. (1997) The Cockayne syndrome B protein, involved in transcription-coupled DNA repair, resides in an RNA polymerase II-containing complex. EMBO J 16: 5955–5965. doi:10.1093/emboj/16.19.5955.

86. Lagerwerf S, Vrouwe MG, Overmeer RM, Fousteri MI, Mullenders LHF (2011) DNA damage response and transcription. DNA Repair (Amst) 10: 743–750. doi:10.1016/j.dnarep.2011.04.024.

87. Brosh RM, Balajee AS, Selzer RR, Sunesen M, Proietti De Santis L, et al. (1999) The ATPase domain but not the acidic region of Cockayne syndrome group B gene product is essential for DNA repair. Mol Biol Cell 10: 3583–3594.

88. Sunesen M, Selzer RR, Brosh RM, Balajee AS, Stevnsner T, et al. (2000) Molecular characterization of an acidic region deletion mutant of Cockayne syndrome group B protein. Nucleic Acids Research 28: 3151–3159.

89. Lake RJ, Geyko A, Hemashettar G, Zhao Y, Fan H-Y (2010) UV-induced association of the CSB remodeling protein with chromatin requires ATP-dependent relief of N-terminal autorepression. Mol Cell 37: 235–246. doi:10.1016/j.molcel.2009.10.027.

90. Lebedev A, Scharffetter-Kochanek K, Iben S (2008) Truncated Cockayne syndrome B protein represses elongation by RNA polymerase I. J Mol Biol 382: 266–274. doi:10.1016/j.jmb.2008.07.018.

91. Dong W, Li Y, Gao M, Hu M, Li X, et al. (2011) IKK  contributes to UVB-induced VEGF expression by regulating AP-1 transactivation. Nucleic Acids Research.

doi:10.1093/nar/gkr1216.

92. Shan Z-X, Lin Q-X, Yang M, Bin Zhang, Zhu J-N, et al. (2011) Transcription factor Ap-1 mediates proangiogenic MIF expression in human endothelial cells exposed to Angiotensin II. Cytokine 53: 35–41. doi:10.1016/j.cyto.2010.09.009.

93. Zenz R, Eferl R, Scheinecker C, Redlich K, Smolen J, et al. (2008) Activator protein 1 (Fos/Jun) functions in inflammatory bone and skin disease. Arthritis Res Ther 10: 201. doi:10.1186/ar2338.

94. Zhang Y, Feng XH, Derynck R (1998) Smad3 and Smad4 cooperate with c-Jun/c-Fos to mediate TGF-beta-induced transcription. Nature 394: 909–913. doi:10.1038/29814.

95. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110: 462–467. doi:10.1159/000084979.

96. Cary LC, Goebel M, Corsaro BG, Wang HG, Rosen E, et al. (1989) Transposon mutagenesis of baculoviruses: analysis of Trichoplusia ni transposon IFP2 insertions within the FP-locus of nuclear polyhedrosis viruses. Virology 172: 156–169.

97. Li X, Lobo N, Bauser CA, Fraser MJ (2001) The minimum internal and external sequence requirements for transposition of the eukaryotic transformation vector piggyBac. Mol Genet Genomics 266: 190–198.

98. Yusa K, Zhou L, Li MA, Bradley A, Craig NL (2011) A hyperactive piggyBac transposase for mammalian applications. Proc Natl Acad Sci USA 108: 1531–1536. doi:10.1073/pnas.1008322108.

99. Whitfield CR, Shilton BH, Haniford DB (2012) Identification of basepairs within Tn5 termini that are critical for H-NS binding to the transpososome and regulation of Tn5 transposition. Mob DNA 3: 7. doi:10.1186/1759-8753-3-7.

100. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456: 53–59. doi:10.1038/nature07517.

101. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10: R25. doi:10.1186/gb-2009-10-3-r25.

102. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, et al. (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol 9: R137. doi:10.1186/gb-2008-9-9-r137.

103. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-Wide Mapping of

in Vivo Protein-DNA Interactions. Science 316: 1497–1502. doi:10.1126/science.1141319.

104. Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, et al. (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods 5: 829–834. doi:10.1038/nmeth.1246.

105. Pepke S, Wold B, Mortazavi A (2009) Computation for ChIP-seq and RNA-seq studies. Nat Methods 6: S22–S32. doi:10.1038/nmeth.1371.

106. Shin H, Liu T, Manrai AK, Liu XS (2009) CEAS: cis-regulatory element annotation system. Bioinformatics 25: 2605–2606. doi:10.1093/bioinformatics/btp479.

107. Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 2: 28–36.

108. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS (2007) Quantifying similarity between motifs. Genome Biol 8: R24. doi:10.1186/gb-2007-8-2-r24.

109. Eferl R, Wagner EF (2003) AP-1: a double-edged sword in tumorigenesis. Nat Rev Cancer 3: 859–868. doi:10.1038/nrc1209.

110. Phillips JE, Corces VG (2009) CTCF: master weaver of the genome. Cell 137: 1194–1211. doi:10.1016/j.cell.2009.06.001.

111. Wagner EF, Eferl R (2005) Fos/AP-1 proteins in bone and the immune system. Immunol Rev 208: 126–140. doi:10.1111/j.0105-2896.2005.00332.x.

112. Chinenov Y, Kerppola TK (2001) Close encounters of many kinds: Fos-Jun interactions that mediate transcription regulatory specificity. Oncogene 20: 2438–2452. doi:10.1038/sj.onc.1204385.

113. Lallemand D, Spyrou G, Yaniv M, Pfarr CM (1997) Variations in Jun and Fos protein expression and AP-1 activity in cycling, resting and stimulated fibroblasts. Oncogene 14: 819–830. doi:10.1038/sj.onc.1200901.

114. Storey J (2002) A direct approach to false discovery rates - Storey - 2002 - Journal of the Royal Statistical Society: Series B (Statistical Methodology) - Wiley Online Library. Journal of the Royal Statistical Society Series B ….

115. Pace JK, Feschotte C (2007) The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. Genome Res 17: 422–432. doi:10.1101/gr.5826307.

116. Mitra R, Fain-Thornton J, Craig NL (2008) piggyBac can bypass DNA synthesis during cut and paste transposition. EMBO J 27: 1097–1109. doi:10.1038/emboj.2008.41.

117. Dreszer TR, Karolchik D, Zweig AS, Hinrichs AS, Raney BJ, et al. (2012) The UCSC Genome Browser database: extensions and updates 2011. Nucleic Acids Research 40: D918–D923. doi:10.1093/nar/gkr1055.

118. Karolchik D (2004) The UCSC Table Browser data retrieval tool. Nucleic Acids Research 32: 493D–496. doi:10.1093/nar/gkh103.

119. Fan H-Y, Trotter KW, Archer TK, Kingston RE (2005) Swapping function of two chromatin remodeling complexes. Mol Cell 17: 805–815. doi:10.1016/j.molcel.2005.02.024.

120. Ouyang X, Li J, Li G, Li B, Chen B, et al. (2011) Genome-wide binding site analysis of FAR-RED ELONGATED HYPOCOTYL3 reveals its novel function in Arabidopsis development. Plant Cell 23: 2514–2535. doi:10.1105/tpc.111.085126.

121. Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonçalves A, et al. (2012) Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. Cell 148: 335–348. doi:10.1016/j.cell.2011.11.058.

122. Lake RJ, Basheer A, Fan H-Y (2011) Reciprocally regulated chromatin association of the Cockayne syndrome protein B and p53. J Biol Chem. doi:10.1074/jbc.M111.252643.

123. Dérijard B, Hibi M, Wu IH, Barrett T, Su B, et al. (1994) JNK1: a protein kinase stimulated by UV light and Ha-Ras that binds and phosphorylates the c-Jun activation domain. Cell 76: 1025–1037.

124. Reno EM, Haughian JM, Jackson TA, Thorne AM, Bradford AP (2009) c-Jun N-terminal kinase regulates apoptosis in endometrial cancer cells. Apoptosis 14: 809–820. doi:10.1007/s10495-009-0354-6.

125. Ip JY, Schmidt D, Pan Q, Ramani AK, Fraser AG, et al. (2011) Global impact of RNA polymerase II elongation inhibition on alternative splicing regulation. Genome Res 21: 390–401. doi:10.1101/gr.111070.110.

126. Hromas R, Wray J, Lee S-H, Martinez L, Farrington J, et al. (2008) The human set and transposase domain protein Metnase interacts with DNA Ligase IV and enhances the efficiency and accuracy of non-homologous end-joining. DNA Repair (Amst) 7: 1927–1937. doi:10.1016/j.dnarep.2008.08.002.

127. De Haro LP, Wray J, Williamson EA, Durant ST, Corwin L, et al. (2010) Metnase promotes restart and repair of stalled and collapsed replication forks. Nucleic Acids Research 38: 5681–5691. doi:10.1093/nar/gkq339.

128. Liu D, Bischerour J, Siddique A, Buisine N, Bigot Y, et al. (2007) The human SETMAR protein preserves most of the activities of the ancestral Hsmar1

transposase. Mol Cell Biol 27: 1125–1132. doi:10.1128/MCB.01899-06.

129. Beck BD, Lee SS, Hromas R, Lee S-H (2010) Regulation of Metnase's TIR binding activity by its binding partner, Pso4. Arch Biochem Biophys 498: 89–94. doi:10.1016/j.abb.2010.04.011.

130. Beck BD, Park S-J, Lee Y-J, Roman Y, Hromas RA, et al. (2008) Human Pso4 is a metnase (SETMAR)-binding partner that regulates metnase function in DNA repair. J Biol Chem 283: 9023–9030. doi:10.1074/jbc.M800150200.

131. Troelstra C, van Gool A, de Wit J, Vermeulen W, Bootsma D, et al. (1992) ERCC6, a member of a subfamily of putative helicases, is involved in Cockayne's syndrome and preferential repair of active genes. Cell 71: 939–953.

132. Bahar B, O'Doherty JV, Maher S, McMorrow J, Sweeney T (2012) Chitooligosaccharide elicits acute inflammatory cytokine response through AP-1 pathway in human intestinal epithelial-like (Caco-2) cells. Molecular Immunology: 1–9. doi:10.1016/j.molimm.2012.03.027.

133. Hipp MS, Urbich C, Mayer P, Wischhusen J, Weller M, et al. (2002) Proteasome inhibition leads to NF-kappaB-independent IL-8 transactivation in human endothelial cells through induction of AP-1. Eur J Immunol 32: 2208–2217. doi:10.1002/1521-4141(200208)32:8<2208::AID-IMMU2208>3.0.CO;2-2.

134. Orjalo AV, Bhaumik D, Gengler BK, Scott GK, Campisi J (2009) Cell surface-bound IL-1alpha is an upstream regulator of the senescence-associated IL-6/IL-8 cytokine network. Proc Natl Acad Sci USA 106: 17031–17036. doi:10.1073/pnas.0905299106.

135. Franceschi C, Bonafè M, Valensin S, Olivieri F, De Luca M, et al. (2000) Inflamm-aging. An evolutionary perspective on immunosenescence. Ann N Y Acad Sci 908: 244–254.

136. Weidenheim KM, Dickson DW, Rapin I (2009) Neuropathology of Cockayne syndrome: Evidence for impaired development, premature aging, and neurodegeneration. Mech Ageing Dev 130: 619–636. doi:10.1016/j.mad.2009.07.006.

137. Kent WJ (2002) BLAT--the BLAST-like alignment tool. Genome Res 12: 656–664. doi:10.1101/gr.229202.

138. Goecks J, Nekrutenko A, Taylor J, Galaxy Team (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11: R86. doi:10.1186/gb-2010-11-8-r86.

139. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, et al. (2005) Galaxy: a platform for interactive large-scale genome analysis. Genome Res 15: 1451–

1455. doi:10.1101/gr.4086505.

140. Blankenberg D, Kuster Von G, Coraor N, Ananda G, Lazarus R, et al. (2010) Galaxy: a web-based genome analysis tool for experimentalists. Curr Protoc Mol Biol Chapter 19: Unit19.10.1–Unit19.10.21. doi:10.1002/0471142727.mb1910s89.

141. Bailey TL, Williams N, Misleh C, Li WW (2006) MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Research 34: W369–W373. doi:10.1093/nar/gkl198.

142. Bradsher J, Auriol J, Proietti-De-Santis L, Iben S, Vonesch JL, et al. (2002) CSB is a component of RNA pol I transcription. Mol Cell 10: 819–829.

143. Abramoff MD, Magalhaes PJ, Ram SJ (2004) Image Processing with ImageJ. Biophotonics International 11: 36–42.

144. Marinescu VD, Kohane IS, Riva A (2005) MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes. BMC Bioinformatics 6: 79. doi:10.1186/1471-2105-6-79.

145. Maglott D (2004) Entrez Gene: gene-centered information at NCBI. Nucleic Acids Research 33: D54–D58. doi:10.1093/nar/gki031.

146. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2011) KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Research 40: D109–D114. doi:10.1093/nar/gkr988.

147. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research 28: 27–30.

148. Steiper ME, Young NM (2006) Primate molecular divergence dates. Molecular Phylogenetics and Evolution 41: 384–394. doi:10.1016/j.ympev.2006.05.021.

149. Chapman JA, Kirkness EF, Simakov O, Hampson SE, Mitros T, et al. (2010) The dynamic genome of Hydra. Nature 464: 592–596. doi:10.1038/nature08830.

150. Dana CE, Glauber KM, Chan TA, Bridge DM, Steele RE (2012) Incorporation of a Horizontally Transferred Gene into an Operon during Cnidarian Evolution. PLoS ONE 7: e31643. doi:10.1371/journal.pone.0031643.g008.

151. Owens JB, Urschitz J, Stoytchev I, Dang NC, Stoytcheva Z, et al. (2012) Chimeric piggyBac transposases for genomic targeting in human cells. Nucleic Acids Research. doi:10.1093/nar/gks309.

152. Davies DR, Goryshin IY, Reznikoff WS, Rayment I (2000) Three-dimensional structure of the Tn5 synaptic complex transposition intermediate. Science 289: 77–85.

153. Hamilton AT, Huntley S, Tran-Gyamfi M, Baggott DM, Gordon L, et al. (2006) Evolutionary expansion and divergence in the ZNF91 subfamily of primate-specific zinc finger genes. Genome Res 16: 584–594. doi:10.1101/gr.4843906.

154. Thomas JH, Schneider S (2011) Coevolution of retroelements and tandem zinc finger genes. Genome Res 21: 1800–1812. doi:10.1101/gr.121749.111.

155. Bryson K, McGuffin LJ, Marsden RL, Ward JJ, Sodhi JS, et al. (2005) Protein structure prediction servers at University College London. Nucleic Acids Research 33: W36–W38. doi:10.1093/nar/gki410.

156. Flicek P, Amode MR, Barrell D, Beal K, Brent S, et al. (2011) Ensembl 2012. Nucleic Acids Research 40: D84–D90. doi:10.1093/nar/gkr991.

157. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410. doi:10.1016/S0022-2836(05)80360-2.

158. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, et al. (2008) NCBI BLAST: a better web interface. Nucleic Acids Research 36: W5–W9. doi:10.1093/nar/gkn201.

159. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947–2948. doi:10.1093/bioinformatics/btm404.

160. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, et al. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology 7: 1–6. doi:10.1038/msb.2011.75.

161. Gouy M, Guindon S, Gascuel O (2010) SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. Mol Biol Evol 27: 221–224. doi:10.1093/molbev/msp259.

162. Felsenstein J (2005) PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author Department of Genome Sciences, University of Washington, Seattle. Available:http://evolution.gs.washington.edu/phylip/. Accessed 28 July 2012.

163. Huson DH, Scornavacca C (2012) Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. Syst Biol. doi:10.1093/sysbio/sys062.

164. Benson DA (2004) GenBank. Nucleic Acids Research 33: D34–D38. doi:10.1093/nar/gki063.

165. Tantin D (1998) RNA polymerase II elongation complexes containing the Cockayne syndrome group B protein interact with a molecular complex containing the transcription factor IIH components xeroderma pigmentosum B

and p62. J Biol Chem 273: 27794–27799.

166. Patel MR, Loo Y-M, Horner SM, Gale M, Malik HS (2012) Convergent Evolution of Escape from Hepaciviral Antagonism in Primates. PLoS Biol 10: e1001282. doi:10.1371/journal.pbio.1001282.t001.

167. Emerman M, Malik HS (2010) Paleovirology—Modern Consequences of Ancient Viruses. PLoS Biol 8: e1000301. doi:10.1371/journal.pbio.1000301.g002.

168. Khan IF, Hirata RK, Russell DW (2011) AAV-mediated gene targeting methods for human cells. Nature Protocols 6: 482–501. doi:10.1038/nprot.2011.301.

169. Mayne LV, Priestley A, James MR, Burke JF (1986) Efficient immortalization and morphological transformation of human fibroblasts by transfection with SV40 DNA linked to a dominant marker. Exp Cell Res 162: 530–538.

170. Sharma GG, Gupta A, Wang H, Scherthan H, Dhar S, et al. (2003) hTERT associates with human telomeres and enhances genomic stability and DNA repair. Oncogene 22: 131–146. doi:10.1038/sj.onc.1206063.

171. Zhang P, Dilley C, Mattson MP (2007) DNA damage responses in neural cells: Focus on the telomere. Neuroscience 145: 1439–1448. doi:10.1016/j.neuroscience.2006.11.052.

172. Indran IR, Hande MP, Pervaiz S (2011) hTERT overexpression alleviates intracellular ROS production, improves mitochondrial function, and inhibits ROS-mediated apoptosis in cancer cells. Cancer Res 71: 266–276. doi:10.1158/0008-5472.CAN-10-1588.

173. Shi W, Hoeflich A, Flaswinkel H, Stojkovic M, Wolf E, et al. (2003) Induction of a senescent-like phenotype does not confer the ability of bovine immortal cells to support the development of nuclear transfer embryos. Biol Reprod 69: 301–309. doi:10.1095/biolreprod.102.012112.

174. Glaser KB, Staver MJ, Waring JF, Stender J, Ulrich RG, et al. (2003) Gene expression profiling of multiple histone deacetylase (HDAC) inhibitors: defining a common gene set produced by HDAC inhibition in T24 and MDA carcinoma cell lines. Mol Cancer Ther 2: 151–163.

175. Mariadason JM, Corner GA, Augenlicht LH (2000) Genetic reprogramming in pathways of colonic cell maturation induced by short chain fatty acids: comparison with trichostatin A, sulindac, and curcumin and implications for chemoprevention of colon cancer. Cancer Res 60: 4561–4572.

176. Burton GR, Nagarajan R, Peterson CA, McGehee RE (2004) Microarray analysis of differentiation-specific gene expression during 3T3-L1 adipogenesis. Gene 329: 167–185. doi:10.1016/j.gene.2003.12.012.

177. Cheng RYS, Zhao A, Alvord WG, Powell DA, Bare RM, et al. (2003) Gene expression dose-response changes in microarrays after exposure of human peripheral lung epithelial cells to nickel(II). Toxicol Appl Pharmacol 191: 22–39.

178. Chiba T, Yokosuka O, Fukai K, Kojima H, Tada M, et al. (2004) Cell growth inhibition and gene expression induced by the histone deacetylase inhibitor, trichostatin A, on human hepatoma cells. Oncology 66: 481–491. doi:10.1159/000079503.

179. Liang G, Gonzales FA, Jones PA, Orntoft TF, Thykjaer T (2002) Analysis of gene induction in human fibroblasts and bladder cancer cells exposed to the methylation inhibitor 5-aza-2'-deoxycytidine. Cancer Res 62: 961–966.

180. Sato N, Maitra A, Fukushima N, van Heek NT, Matsubayashi H, et al. (2003) Frequent hypomethylation of multiple genes overexpressed in pancreatic ductal adenocarcinoma. Cancer Res 63: 4158–4166.

181. Nuytten M, Beke L, Van Eynde A, Ceulemans H, Beullens M, et al. (2008) The transcriptional repressor NIPP1 is an essential player in EZH2-mediated gene silencing. Oncogene 27: 1449–1460. doi:10.1038/sj.onc.1210774.

182. Van Dessel N, Beke L, Görnemann J, Minnebo N, Beullens M, et al. (2010) The phosphatase interactor NIPP1 regulates the occupancy of the histone methyltransferase EZH2 at Polycomb targets. Nucleic Acids Research 38: 7500–7512. doi:10.1093/nar/gkq643.

183. Crea F, Paolicchi E, Marquez VE, Danesi R (2011) Polycomb genes and cancer: Time for clinical application? Crit Rev Oncol Hematol. doi:10.1016/j.critrevonc.2011.10.007.

184. Lipnik K, Naschberger E, Gonin-Laurent N, Kodajova P, Petznek H, et al. (2010) Interferon gamma-induced human guanylate binding protein 1 inhibits mammary tumor growth in mice. Mol Med 16: 177–187. doi:10.2119/molmed.2009.00172.

185. Durfee LA, Huibregtse JM (2010) Identification and Validation of ISG15 Target Proteins. Subcell Biochem 54: 228–237. doi:10.1007/978-1-4419-6676-6_18.

186. Broquet AH, Hirata Y, McAllister CS, Kagnoff MF (2011) RIG-I/MDA5/MAVS are required to signal a protective IFN response in rotavirus-infected intestinal epithelium. The Journal of Immunology 186: 1618–1626. doi:10.4049/jimmunol.1002862.

187. Daffis S, Szretter KJ, Schriewer J, Li J, Youn S, et al. (2010) 2'-O methylation of the viral mRNA cap evades host restriction by IFIT family members. Nature 468: 452–456. doi:10.1038/nature09489.

188. Huang I-C, Bailey CC, Weyer JL, Radoshitzky SR, Becker MM, et al. (2011) Distinct patterns of IFITM-mediated restriction of filoviruses, SARS coronavirus,

and influenza A virus. PLoS Pathog 7: e1001258. doi:10.1371/journal.ppat.1001258.

189. Watson IR, Irwin MS, Ohh M (2011) NEDD8 pathways in cancer, Sine Quibus Non. Cancer Cell 19: 168–176. doi:10.1016/j.ccr.2011.01.002.

190. Nagai MA, Fregnani JHTG, Netto MM, Brentani MM, Soares FA (2007) Down-regulation of PHLDA1 gene expression is associated with breast cancer progression. Breast Cancer Res Treat 106: 49–56. doi:10.1007/s10549-006-9475-6.

191. Song G, Fleming J-AGW, Kim J, Spencer TE, Bazer FW (2011) Pregnancy and interferon tau regulate DDX58 and PLSCR1 in the ovine uterus during the peri-implantation period. Reproduction 141: 127–138. doi:10.1530/REP-10-0348.

192. Lowman XH, McDonnell MA, Kosloske A, Odumade OA, Jenness C, et al. (2010) The proapoptotic function of Noxa in human leukemia cells is regulated by the kinase Cdk5 and by glucose. Mol Cell 40: 823–833. doi:10.1016/j.molcel.2010.11.035.

193. Chibi M, Meyer M, Skepu A, G Rees DJ, Moolman-Smook JC, et al. (2008) RBBP6 interacts with multifunctional protein YB-1 through its RING finger domain, leading to ubiquitination and proteosomal degradation of YB-1. J Mol Biol 384: 908–916. doi:10.1016/j.jmb.2008.09.060.

194. Dorniak P, Bazer FW, Spencer TE (2011) Prostaglandins regulate conceptus elongation and mediate effects of interferon tau on the ovine uterine endometrium. Biol Reprod 84: 1119–1127. doi:10.1095/biolreprod.110.089979.

195. Stewart MD, Stewart DM, Johnson GA, Vyhlidal CA, Burghardt RC, et al. (2001) Interferon-tau activates multiple signal transducer and activator of transcription proteins and has complex effects on interferon-responsive gene transcription in ovine endometrial epithelial cells. Endocrinology 142: 98–107.

196. Liu J, Wennier S, Zhang L, McFadden G (2011) M062 is a host range factor essential for myxoma virus pathogenesis and functions as an antagonist of host SAMD9 in human cells. J Virol 85: 3270–3282. doi:10.1128/JVI.02243-10.

197. Liu X, Lei X, Zhou Z, Sun Z, Xue Q, et al. (2011) Enterovirus 71 induces degradation of TRIM38, a potential E3 ubiquitin ligase. Virol J 8: 61. doi:10.1186/1743-422X-8-61.

198. Pertel T, Hausmann S, Morger D, Züger S, Guerra J, et al. (2011) TRIM5 is an innate immune sensor for the retrovirus capsid lattice. Nature 472: 361–365. doi:10.1038/nature09976.

# Appendix 1: Consequences of microarray database bias

### Comparison of the CS1AN and UVSS1KO datasets using L2L and MSigDB revealed database bias

We were puzzled by the apparent discrepancy between our analysis of the UVSS1KO microarray data, and our previous analysis of microarray data for CS1AN cells [6]. CS1AN is a CS patient-derived compound heterozygote which continues to express intact CSB-PGBD3 fusion protein [14] along with severely truncated CSB polypeptides [131]; UVSS1KO is a patient-derived CSB-null line in which a homozygous nonsense mutation at codon 77 abolishes expression of both the CSB-PGBD3 fusion protein [14] and stable CSB fragments [15]. Using L2L for data analysis [39], addition of CSB to CS1AN (which should restore the normal genotype) exhibited a strong chromatin remodeling signature but did not appear to affect interferon-related genes [6]. In contrast, addition of CSB to UVSS1KO-derived cells expressing the CSB-PGBD3 fusion protein (which should also restore the normal genotype) damped the interferon-like response induced by expression of CSB-PGBD3 fusion protein alone, but exhibited no obvious chromatin remodeling signature (Table 2-S1A).

We initially wondered whether the method of immortalization might account for the discrepancy: UVSS1KO fibroblasts are SV40-transformed [15] in contrast to our CS1AN fibroblast line which we rederived from primary CS1AN cells (Coriell Institute Repository, Camden NJ) by transformation with hTERT [6] instead of using the original SV40-transformed CS1ANSV line that has been propagated for 25 years [169]. Unlike transformation with SV40, transformation with hTERT is capable of enhancing both DNA repair and genomic stability [170,171] through two distinct mechanisms: one subpopulation of hTERT associates directly with telomeres and appears to upregulate DNA damage repair pathways, while another subpopulation of hTERT associates with mitochondria where it generates a reducing environment within the organelle and potentiates disposal of ROS [172]. However, the discrepancy between the UVSS1KO and CS1AN datasets is unlikely to reflect the mode of transformation because the differences in gene expression between the two cell lines do not exhibit any trace of a DNA repair signature and only ambiguous redox indicators (Table 2-S1A).

We next asked whether the microarray databases themselves might account for the apparent discrepancy. We originally analyzed the CS1AN and UVSS1KO datasets using our L2L Microarray Software Suite and Database [39] although we were aware of the MSigDB software and database (www.broadinstitute.org/gsea/msigdb/index.jsp and [83]). The L2L and MSigDB suites are functionally similar, but upon closer examination we discovered that the curated MSigDB database currently includes many interferon-

related lists that were not available in 2005, but surprisingly *lacks* many of chromatin remodeling lists that *were* available in 2005 and were included in the L2L database. Specifically, of the 23 gene sets cited in Figure 1 of Newman et al. [14] as responsible for the chromatin-remodeling signature, only 5 were present in MSigDB. All datasets from 4 studies [173-176] were absent; and although 3 other studies [177-180] were included in the MSigDB database, none of these were in the top 50 datasets for genes up- or down-regulated by >2-fold when the CS1AN datasets were reanalyzed with MSigDB as described below (L.T. Gray, data not shown).

Thus database bias explains why MSigDB readily detected the strong interferon signature for expression of the CSB-PGBD3 fusion protein in CSB-null UVSS1KO cells (Table 2-S4A) but failed to identify the chromatin remodeling signature detected by L2L (Tables 2-S1A and 2-S2) when CSB is coexpressed with endogenous CSB-PGBD3 fusion protein in CS1AN cells [6]. Our concern that the databases themselves could influence, or even obscure, significant functional signatures in microarray datasets was confirmed when we noticed that MSigDB tends to find prominent cancer signatures as well as hypoxia and extracellular matrix signatures associated with invasion and metastasis (Table 2-S4A). We conclude that evolving microarray databases often reflect current research priorities (in this case cancers) and may bias the top hits toward the most highly represented research areas.

**Are weaker or related signatures obscured by the strong interferon-like response?**

The realization that MSigDB tends to identify cancer-related signatures (above) led us to ask whether the cancer, chromatin remodeling, and inflammation signatures induced by the CSB-PGBD3 fusion protein in UVSS1KO cells are independent signatures or aspects of the interferon response. To do this, we instructed MSigDB to identify the 50 datasets in the MSigDB database that best matched genes upregulated 2-fold or more by expression of the CSB-PGBD3 fusion protein in the CSB-null UVSS1KO pool (Table 2-S1). Of the top 50 datasets (Table 2-S4A, *upper panel*), 47 could be sorted into 4 broad categories: (6 chromatin, 19 cancer, 15 interferon response, and 7 inflammation). The cancer and interferon response datasets dominated, but the 4 categories were interspersed with each other over the full range of p-values from $<10^{-16}$ to $10^{-11}$. MSigDB also ranked the 260 genes or loci that were responsible for these overlaps based on the frequency of occurence in all overlaps (Table 2-S4A, *lower panel*). The datasets were then separated into the 4 categories above, and the frequency with which each gene appeared in each category was tabulated using a Perl script and sorted in Excel (Table 2-S4B). Remarkably, the rank order of the genes in each of the 4 categories was similar, strongly suggesting that the chromatin, cancer, and

inflammation signatures are not independent functional consequences of the CSB-PGBD3 fusion protein but instead reflect the presence of a common interferon-related component (Table 2-S4B).

The role of the interferon response in tumorigenesis is complex and poorly understood, but it is clear that the response differs greatly from one cell type to another, and can either antagonize or promote tumor formation or apoptosis depending on the type and history of the tumor [44]. Thus it is not surprising that the cancer-related datasets in MSigDB (Table 2-S4A, *upper panel*) and the genes responsible for overlap with the UVSS1KO dataset (Table 2-S4A, *lower panel*) reflect a broad range of cancers as well as interferon-related (bold red) and/or innate immunity components (highlighted in blue). Similarly, the most prominent chromatin remodeling dataset in Table S4A [181] reflects the ability of the phosphatase interactor NIPP1 to target the repressive Polycomb complex PRC2 which is upregulated in many cancers, and associated with metastasis [182,183].

# Appendix 2: Vignettes of the 15 IFN-related genes upregulated by expression of CSB in the CSB compound heterozygote CS1AN line.

• DDX58 aka RIG-I is an RNA helicase that recognizes viral RNA and participates in a positive feedback loop with STAT1 [49].
• GBP1 is a Guanine Binding Protein of the dynamin family usually involved in vesicle scission and trafficking [184].
• The core E1, E2 and E3 enzymes for conjugation of the human interferon-induced ubiquitin- like antiviral protein ISG15 are Ube1L, UbcH8 and Herc5, all of which are transcriptionally induced by Type 1 interferon [185].
• IFIH1 is an interferon-induced RNA helicase aka MDA5, and like RIG-I combines the RNA helicase domain with N-terminal caspase activation and recruitment domains or CARDs [186].
• IFITs are IFN-induced proteins with tetratricopeptide repeats aka interferon-stimulated genes (ISGs) implicated in regulation of translation [187].
• IFITM1, 2, and 3 are interferon-inducible transmembrane proteins 1, 2, and 3 that inhibit influenza A virus infection mediated by the hemagglutinin protein [188].
• ISG15 is 1 of 17 ubiquitin-like proteins (UBLs) from 9 phylogenetically distinct classes (NEDD8, SUMO, ISG15, FUB1, FAT10, Atg8, Atg12, Urm1, and UFM1) that are conjugated to proteins [189].
• PHLDA1 is a pleckstrin homology-like domain protein aka TDAG51 [190].
• PLSCR1 is phospholipid scramblase 1, an interferon-inducible protein that inserts into the plasma membrane or binds DNA in the nucleus depending on its state of palmitoylation [56]. PLSCR1 and the RIG-I/DDX58 DEAD box helicase are both induced by pregnancy- associated IFNô through the STAT1-dependent pathway [191].
• PMAIP1 aka Noxa is a proapoptotic protein [192].
• RBBP6 is a RING finger protein that participates in ubiquitination and proteosomal degradation
of YB-1 [193].
• RSAD2 is a pregnancy-related gene stimulated by IFNô [194] which is secreted by the trophoblast in sheep and cows, and activates multiple STATs in endometrial epithelial cells [195].
• SAMD9 is an interferon-regulated sterile alpha motif domain protein implicated in innate immunity to poxviruses and in inflammatory disorders [196].
• TRIM or tripartite motif proteins, which have three zinc-binding domains, a RING domain, type 1 and type 2 B-boxes, and a coiled-coil region, are thought to be involved in protein ubiquitination [197]. For example, TRIM5 is an innate immune sensor for retrovirus capsid lattices; TRIM5 forms an hexagonal lattice which interacts with the retrovirus capsid lattice, enhancing TRIM5 E3 activity and ultimately inducing transcription of AP-1 and NF-kB- dependent factors [198].

# Appendix 3: Perl Scripts

The following are key Perl scripts that were used to analyze the ChIP-seq data presented in Chapter 3.

sam_sorted_to_fragment_wig.pl was used to convert location sorted paired end read location assignments to fragment overlap profiles in wiggle format for visualization of fragment density, analysis of peak summit location, and pileups of many regions.

wig_peak_summits.pl was used to examine the fragment overlaps generated by sam_sorted_to_fragment_wig.pl at locations of ChIP-seq tag enrichment located using peak finding algorithms MACs, QuEST, or ERANGE. This script is very flexible, and can be used to find the highest value in a WIG file for each region provided in BED format.

wig_pileup_bed.pl was used to generate graphs for visualization of WIG score density for either single regions (as for PGBD3 and PGBD3 pseudogenes in Chapter 3), or for the cumulative scores over many regions (as for MER85s, TRE, TEAD1, and CTCF peaks in Chapter 3). This script calculates the fragment overlap relative to the 5' end of each region in a BED file provided as input. If multiple regions are provided, the sum of the overlaps at each position relative to the 5' end of each provided region are reported.

sam_insert_size.pl was used to calculate the average and standard deviation of the fragment lengths in the paired-end read location files. This calculation was required for submission to GEO.

## sam_sorted_to_fragment_wig.pl

```
#!/usr/bin/perl
# ---------------
# sam_sorted_to_fragment_wig.pl - converts a sorted, paired-end SAM file to a
# variable-step WIG file of fragment
# overlap counts. Requires the read length of the raw reads used to generate the SAM
# file, as well as
# the minimum number of overlapping fragments you want displayed. The latter can
# greatly reduce the file
# size by leaving out orphaned fragments while retaining regions that are
# significantly enriched. Use 0
# for min_fragments to keep all reads.
# ---------------
# usage: sam_sorted_to_fragment_wig.pl sam_file wig_output read_length min_fragments
# example: ./sam_sorted_to_fragment_wig.pl csb-pgbd3_paired.sorted.sam
# csb-pgbd3_paired.wig 36 5
# This will create a WIG file from data generated with 36 bp reads from each end of
# sequenced fragments
# and will output only regions with at least 5 overlapping fragments.
# ---------------
# SAM files must be from paired-end data, and can be generated from raw reads using
# the read alignment program Bowtie (http://bowtie-bio.sourceforge.net/).
# SAM files must be sorted before this script is used. SAM files can be sorted using
# IGVTools (http://www.broadinstitute.org/igv/igvtools).
# For more information about the SAM file format, see the SAMtools page
# (http://samtools.sourceforge.net/).
# For a description of WIG format, see the UCSC Genome Browser description of wiggle
# format (https://cgwb.nci.nih.gov/goldenPath/help/wiggle.html).
# ---------------

if(@ARGV < 3) {

	print "usage: sam_sorted_to_fragment_wig.pl sam_file wig_output "
	. "read_length min_fragments\n";

} else {

	$sam_in = @ARGV[0];
	$wig_out = @ARGV[1];
	$read_length = @ARGV[2] - 1;
	$min_frags = @ARGV[3];

	open(SAM, "<$sam_in");
	open(WIG, ">$wig_out");
	close(WIG);

	$last_end = 0;
	@cur_peak = ();
	$cur_chr = "chrZ";

	#subroutine for calculating the overlaps given a cur_peak array
	sub calc_overlaps {
		#only process if there are at least the minimum number of fragments in
		# the peak
		if(@cur_peak >= $min_frags) {
			#the start of the peak will be the start of the first fragment in
			#the array
			($cur_start, $z) = split(/\t/, @cur_peak[0]);

			#to get the end, we have to look for the largest value in the
			# fragment ends
			$cur_end = 0;
```

```
		#cycle through all the fragments in the array
		foreach(@cur_peak) {
			#get the value for the 3' position of the fragment
			($z ,$test_end) = split(/\t/, $_);
			#if it's bigger than what's stored as $cur_end, update
			# $cur_end with the larger value
			if($test_end > $cur_end) {
				$cur_end = $test_end;
			}
		}

		#get the length of the current peak
		$length = $cur_end - $cur_start;

		#fill an array of the peak length with 0's as a baseline
		for($x = 0; $x < $length; $x++) {
			push(@map, 0);
		}

		#look through each fragment and increment the map at each position
		# the fragment overlaps
		for($j = 0; $j < @cur_peak; $j++) {
			#get the start and end of the current fragment
			($frag_start, $frag_end) = split(/\t/, @cur_peak[$j]);
			#adjust the start and end values so that the first position
			# is 0, as in the array
			$map_start = $frag_start - $cur_start;
			$map_end = $frag_end - $cur_start;
			#for each position in the map array that the fragment
			#overlaps,
			for($k = $map_start; $k < $map_end; $k++) {
				#increment the value of the map position
				$old_height = @map[$k];
				$map[$k] = $old_height + 1;
			}
		}

		open(WIG, ">>$wig_out");
		#output the overlaps at each position in the map array
		for($x = 0; $x < @map; $x++) {
			#readjust the relative position in the map to the actual
			# genomic position
			$position = $x + $cur_start;
			#print the position and overlap to the output
			print WIG $position . "\t" . @map[$x] . "\n";
		}
		close(WIG);

		@map = ();

	}
}

while(<SAM>) {

	$line = $_;
	chomp($line);
	@line = split(/\t/,$line);

	#Make sure this isn't a header line or from chromosome M
	if ( @line[0] eq "\@HD" || @line[0] eq "\@SQ" || @line[0] eq "\@PG"
	|| @line[2] eq "chrM") {
	} elsif ( @line[1] == 99 || @line[1] == 163) {
```

```
		#if not, only process lines with flags 99 or 163.

			#get information about the fragment position
			$start = @line[3];
			#have to add the read length to the 3' end, because SAM format
			# uses only the 5'-most position
			$end = @line[7] + $read_length;
			#this is the fragment line as it will be stored in the array for
			# retrieval:
			$fragment = $start . "\t" . $end;

			#Check to see if this is a new chromosome
			if ( @line[2] ne $cur_chr ) {
				#if so, output the final peak from the end of the last
				# chromosome

				calc_overlaps();

				#and start processing the new chromosome.
				$cur_chr = @line[2];
				print "Now processing " . $cur_chr . "\n";

				#output the header line for the new chromosome to the wig
				# file
				open(WIG, ">>$wig_out");
				print WIG "variableStep chrom=" . $cur_chr . "\n";
				close(WIG);

				#reset the end point position and current peak array
				$last_end = 0;
				@cur_peak = ();

			} else {
				#if this fragment is outside the last peak, output the last
				# peak and start storing
				#fragments for a new peak.
				if($start > $last_end) {

					calc_overlaps();

					@cur_peak = ();
				}

			}

			push(@cur_peak, $fragment);

			if($end > $last_end) {
				$last_end = $end;
			}
		}
	}

	#output the last peak

	calc_overlaps();

	close(SAM);

}
```

## wig_peak_summits.pl

```
#!/usr/bin/perl
# wig_peak_summits.pl - looks through a wiggle file for the highest points in a given
# set of bed regions.
# These are the areas of highest fragment overlap - height, start, end, length, and
# center are calculated and output.
# usage: wig_peak_summits.pl in_wig in_peaks out_summits

if (@ARGV != 3) {
	print "usage: wig_peak_summits.pl in_wig in_peaks out_summits\n";
} else {

	$wig_in = @ARGV[0];
	$peaks_in = @ARGV[1];
	$out_file = @ARGV[2];

	open(PEAKS, "<$peaks_in");
	while(<PEAKS>) {

		$line = $_;
		chomp($line);

		@line_split = split(/\t/, $line);

		push(@{"@line_split[0]"}, $line);

	}
	close(PEAKS);

	sub process_chr {
		print "Finished reading $cur_chr.\n";
			#if it is, check through the regions for that chromosome to do the
			# calculations.

			for ($i = 0; $i < @{"$cur_chr"}; $i++) {

				#split up the region. 0 is chr, 1 is start, 2 is end.
				@region_split = split(/\t/, @{"$cur_chr"}[$i]);

				#set the variables to keep track of
				$high_start = 0;
				$high_end = 0;
				$high_value = 0;
				$find_end = 0;

				#look through each wiggle point for the chromosome
				for ($j = 0; $j < @chr_wig; $j++) {

					#split the wiggle line. 0 is position, 1 is height
					@split_wig = split(/\t/, @chr_wig[$j]);

					if ( @split_wig[0] < @region_split[1] ) {
						#if the wiggle position is below the start of
						# the current region
						#get rid of the wiggle position so the search
						# is faster
						shift(@chr_wig);
						$j--;
					} elsif (@split_wig[0] > @region_split[2]) {
						#if the wiggle position is above the end of
						# the current region,
						#stop looking through the wiggle data
```

```
                                last;

                        } elsif ( @split_wig[0] <= @region_split[2] ) {
                                #if it does fall in the region, check for top
                                # position

                                if (@split_wig[1] > $high_value) {
                                        #if this is a new high point, set the
                                        # location, the height, and
                                        #set the variable to check for where it
                                        # ends
                                        $high_start = @split_wig[0];
                                        $high_value = @split_wig[1];
                                        $find_end = 1;

                                } elsif ( ($find_end == 1) &&
                                (@split_wig[1] < $high_value) ) {
                                        #if a new high point has been found,
                                        # and the current location is
                                        #lower than the high point, it's the
                                        # end of the overlap plateau.
                                        $high_end = @split_wig[0] - 1;
                                        $find_end = 0;

                                }

                        }

                }

        #in some cases, the highest overlap for the peak is at the end of
        # the region.
        #when that happens, the end won't be found properly, and should be
        # the end of the region.

        if ($high_end < $high_start) {
                $high_end = @region_split[2];
        }

        #after looking through the wiggle positions, print the top point
        # for that peak.

        $summit_length = $high_end - $high_start + 1;
        $center = int($high_start + $summit_length / 2);
        $rel_center = $center - @region_split[1];
        $peak_length = @region_split[2] - $region_split[1];
        $rel_center_ratio = $rel_center / $peak_length;

        print OUTPUT @{"$cur_chr"}[$i] . "\t" . $high_start . "\t" .
        $high_end . "\t" . $high_value . "\t" . $summit_length . "\t" .
        $center . "\t" . $rel_center . "\t" . $rel_center_ratio . "\n";

        }

        @chr_wig = ();
        $cur_chr = @chr_check[1];

}

#Then, look through the wiggle file for positions that are in the regions
$display_counter = 0;
$total_counter = 0;
```

```
	$cur_chr = "Regions";

	open(WIG, "<$wig_in");
	open(OUTPUT, ">$out_file");

	#Read in the wiggle file. I'll do this a chromosome at a time, then process
	# that chromosome.
	while(<WIG>) {

		$line = $_;
		chomp($line);

		#check to see if the line is a chromosome header
		@chr_check = split(/=/,$line);

		if (@chr_check[0] eq "variableStep chrom") {
			#if this is a new chromosome, run the process chromosome
			# subroutine, above.
			process_chr();

		} else {

			push(@chr_wig, $line);

		}

		$display_counter++;
		$total_counter++;

		if ($display_counter == 10000000) {
			print "Processed $total_counter wiggle lines\n";
			$display_counter = 0;
		}

	}

	#we'll need to run the process_chr one more time in order to get the last
	# chromosome output:
	process_chr();

	close(WIG);
	close(OUTPUT);


}
```

## wig_pileup_bed.pl

```
#!/usr/bin/perl
# wig_pileup_bed.pl - Reads a wiggle file to get the cumulative fragment pileup of
# reads over a set of regions defined by a start site, a direction, and the number of
# bases up and downstream from the start sites that should be analyzed.
# usage: wig_pileup_bed.pl in_locations in_wig upstream downstream out_file

if (@ARGV != 5) {
	print "usage: wig_pileup_bed.pl in_locations in_wig upstream downstream
out_file \n";
} else {

	#declare input variables
	$in_loc = @ARGV[0];
	$in_wig = @ARGV[1];
	$up = @ARGV[2];
	$down = @ARGV[3];
	$out_file = @ARGV[4];

	#Generate an array to store the pileup values in, starting out at zero for each
	# position
	$total_region = $up + $down;

	for($i = 0; $i < $total_region; $i++) {
		push(@pileup, 0);
	}

	#calculate the start and end positions for each location to make searching
	# through the
	#wiggle file simpler later

	open(LOC, "<$in_loc");
	while(<LOC>) {

		$line = $_;
		chomp($line);

		@line_split = split(/\t/, $line);

		if(@line_split[5] eq "+") {
			$start = @line_split[1] - $up;
			$end = @line_split[1] + $down;
		} else {
			$start = @line_split[2] - $down;
			$end = @line_split[2] + $up;
		}

		$new_line = $start . "\t" . $end . "\t" . @line_split[5];

		#put the adjusted locations in arrays for each chromosome
		push(@{"@line_split[0]"}, $new_line);

	}
	close(LOC);

	#then parse through the wiggle file to check if each line falls within a region
	# of interest
	$display_counter = 0;
	$total_counter = 0;
	$cur_chr = "Locations";
	open(WIG, "<$in_wig");
	while(<WIG>) {
```

```
$line = $_;
chomp($line);

#check to see if the line is a chromosome header
@chr_check = split(/=/,$line);

if (@chr_check[0] eq "variableStep chrom") {
	print "Finished processing $cur_chr.\n";
	$cur_chr = @chr_check[1];

} else {

	#if it's not a header, check through the regions for the current
	# chromosome

	for ($i = 0; $i < @{"$cur_chr"}; $i++) {

		@region_split = split(/\t/, @{"$cur_chr"}[$i]);

		#check to see if the wiggle location falls in the region

		@split_wig = split(/\t/, $line);

		if ( @split_wig[0] < @region_split[0] ) {
			#if the wiggle position is below the start of the
			# current region stop looping through regions. This
			# should improve speed.
			last;
		} elsif (@split_wig[0] > @region_split[1]) {
			#if the wiggle position is above the end of the
			# current region, the region should be removed from
			# the array to speed comparisons
			shift(@{"$cur_chr"});
			$i--;
		} elsif ( @split_wig[0] <= @region_split[1] ) {
			#if it does fall in the region, figure out what
			# position it's in this will be dependent on the
			# direction of the region (@region_split[3])

			if (@region_split[2] eq "+") {
				$corrected_position = @split_wig[0] -
				@region_split[0];
			} elsif (@region_split[2] eq "-") {
				$corrected_position = @region_split[1] -
				@split_wig[0];
			}

			#finally, increase the value of the corrected
			# position by the height value for the position from
			# the wiggle file.
			$new_height = @pileup[$corrected_position] +
			@split_wig[1];
			splice(@pileup,$corrected_position,1,$new_height);

		}

	}

}

$display_counter++;
$total_counter++;
```

```
		if ($display_counter == 10000000) {
			print "Processed $total_counter wiggle lines\n";
			$display_counter = 0
		}

	}

	#After going through the wiggle file, we should be able to output the overlaps

	open(OUTPUT, ">$out_file");

	for ($i = 0; $i < @pileup; $i++) {

		$position = 0 - $up + $i;

		print OUTPUT $position . "\t" . @pileup[$i] . "\n";

	}

}
```

## sam_insert_size.pl

```
#!/usr/bin/perl
# sam_insert_size.pl - calculates the average and stdev of insert sizes in a SAM file.
# usage: sam_insert_size.pl sam_in

if(@ARGV != 1) {

	print "usage: sam_insert_size.pl sam_in\n";

} else {

	$in_file = @ARGV[0];
	$length = @ARGV[1];

	open(INPUT, "<$in_file");

	$total_length = 0;
	$fragment_count = 0;
	@fragment = ();

	while(<INPUT>) {

		$line = $_;
		chomp($line);

		@line_split = split(/\t/,$line);

		#Make sure this isn't a header line or from chromosome M
		if ( @line_split[0] eq "\@HD" || @line_split[0] eq "\@SQ" ||
		@line_split[0] eq "\@PG" || @line_split[2] eq "chrM") {
		} elsif ( @line_split[1] == 99 || @line_split[1] == 163) {
		#if not, only process lines with flags 99 or 163.

			$fragment_length = @line_split[8];
			$total_length += $fragment_length;
			push(@fragment, $fragment_length);
			$fragment_count++;

		}

	}

	close(INPUT);

	$average = $total_length / $fragment_count;

	$top = 0;

	foreach(@fragment) {

		$top += ( $_ - $average )**2;

	}

	$stdev = sqrt( $top / $fragment_count );

	print "average: " . $average . "\n";
	print "stdev: " . $stdev . "\n";

}
```

# VITA

Lucas Gray was born in Forks, Washington in 1984. He grew up in Centralia, Washington, where he first learned molecular cloning techniques at Centralia High School in the classroom of Mike Stratton. After graduation from CHS in 2003, Lucas went to Washington State University, where he earned a Bachelor of Science in Biochemistry with a minor in Molecular Biology in 2006. While at WSU, he worked in the laboratory of Dr. Norman Lewis, where he cloned several putative peroxidase genes used for plant secondary metabolism and lignin formation in Arabidopsis. In 2006, he moved to Seattle, where he joined the lab of Dr. Alan Weiner to investigate the function and conservation of the human CSB-PGBD3 fusion protein. He earned his Doctor of Philosophy in Biochemistry at the University of Washington in 2012.