BILKE ET AL.

SUPPLEMENTAL INFORMATION

Detailed information for ChIP-seq results

The input-tags were, in general, distributed evenly. However, several regions showed up to two-fold increased read density compared to the theoretical flat background. Those regions were frequently located close to transcription start sites, most likely due to preferential shearing at regions of open chromatin (**Supplemental Figure SF1 (a)**). As a consequence of this observation, the input DNA tags were used as a background in the statistical analysis of the ChIP data to reduce the risk of false positive detection of binding sites. On a per chromosome scale, the relative number of reads for the ChIP-selected DNA was close to that obtained for unselected input DNA (**Supplemental Table ST2)**, with the exception of the gene rich chromosome 19, where the relative number of reads was 17% and 30% increased over input for EWSR1/FLI1 and E2F3, respectively. On a local scale, the ChIP data frequently showed regions of strongly increased read density, up to several 100 fold over input.

For validation of EWSR1/FLI1 ChIP-seq sensitivity and specificity, we tested for signals in the vicinity of previously established direct EWSR1/FLI1 candidate target genes *TGFBR2* (Hahm et al. 1999), *Id2* (Fukuma et al. 2003*)*, *CDKN1A* (Nakatani et al. 2003), IGFBP3 (Prieur et al. 2004)*, TERT* (Fuchs et al. 2004), *STYXL1* (Siligan et al. 2005), *PTPL1* (Abaan et al. 2005), *CAV1* (Tirado et al. 2006), *NR0B1* (Kinsey et al. 2006), *PLD2* (Kikuchi et al. 2007), *GLI1* (Beauchamp et al. 2009), *GSTM4* (Luo et al. 2009), NKX2.2 (Smith et al. 2006). Pronounced ChIP signals in the vicinity of these genes were observed with the exception of *TERT* and *PTPL1* (for representative examples**, Supplemental Figure SF1 (b)).** In these genes, EWSR1/FLI1 binding was found close to the transcription start site, further upstream and/or in introns. A weak signal adjacent to the previously described EWSR1/FLI1 binding $(GGAA)_n$ microsatellite 1.4kb upstream of *NR0B1* (Gangwal and Lessnick 2008) did not achieve the significance threshold.

Identification of ERG binding regions in prostate cells.

Aligned ERG ChIP-seq reads were downloaded from GEO (GSE14092). Significantly

enriched binding regions were identified using identical algorithm and parameters used for identification of E2F3, EWSR1/FLI-1 binding regions, with the exception of the background model where as a substitute for unselected reads, which were not measured in the experiment, a flat Poisson distribution was used. Regions were defined as ”overlapping” when they occupied at least a single common nucleotide.

Comparison of gene expression changes in VCaP and A673 cells

Up- and Down-regulated genes in A673 cells were derived from the genes associated with the principal components PCA1 and PCA2 by merging the corresponding data sets PCA1+, PCA2+ and PCA1-, PCA2- with $|r| > 0.8$, respectively. For VCaP, the preprocessed series gene expression matrix files described in GSE16671 were downloaded from GEO. Array probes were ranked according to their t-statistics (ERG knockdown vs. scramble shRNA) and the top 1000 unique up- and down- regulated gene symbols were selected as knockdown responsive genes. The overlap of gene sets was calculated based on unique gene symbols. From the size of these sets, p-values were calculated based on the cumulative hyper-geometric distribution.

$(GGAA)_n$ microsatellites in EWSR1/FLI1 binding regions

While we confirm 93% (228) of the genomic binding regions for EWSR1/FLI1 described in a previous investigation (Guillon et al. 2009) our study identified ten times more uniquely aligned sequence tags than previously reported. Thus the identification of proximal EWSR1/FLI1 binding regions, which we found to represent less than 20% of EWSR1/FLI1 binding events, may be related to the much higher sequencing depth achieved in our study. On the other hand, it has been demonstrated that distant binding, which may occur at a distance of up to several megabases, frequently occupies $(GGAA)_n$ microsatellites representing highly repetitive arrays of ETS recognition core motifs (Gangwal and Lessnick 2008; Guillon et al. 2009). In-vitro evidence suggests that EWSR1/FLI1 affinity to such sites increases with GGAA copy number (Gangwal and Lessnick 2008; Guillon et al. 2009). Thus, such microsatellites may constitute high affinity genomic binding regions for EWSR1/FLI1 possibly explaining their prevalence in ChIP-seq with lower sequencing depth. The contribution of $(GGAA)_n$ microsatellites to the EWSR1/FLI1 binding spectrum cannot be fully appreciated in a genomic sequencing study due to

the inability to align repetitive sequence tags to unique genomic regions. Consequently, while we validated our ChIP-seq results by identifying EWSR1/FLI1 binding in 10 of 12 previously established direct targets, neither the Guillon study nor our study were able to identify EWSR1/FLI1 bound to the prototype target gene assumed to be activated by EWSR1/FLI1 through a $(GGAA)_n$ microsatellite, *NR0B1 (Gangwal and Lessnick 2008)*. A small peak below the threshold of significance was observed immediately flanking the $(GGAA)_n$ microsatellite 1.4 kb upstream of *NR0B1*, possibly identifying the margin of the EWSR1/FLI1 binding region. However in the unaligned EWSR1/FLI1 ChIP sequences, we found that the frequency of the $(GGAA)_n$ motifs was approximately four times higher compared to input DNA, in line with the previously suggested preference of EWSR1/FLI1 for such genomic regions. Furthermore, comparison of the binding regions detected in this study to the microsatellite binding regions previously described[14] found an over 90% overlap of both the reported high- (95 of 104 sequences with 3 or more GGAA repeats) and lower-repeat regions (130 of 141). In comparison, randomization provided an estimated overlap of approximately only one matched region per dataset. Performing the same analysis for the E2F3 regions found no overlap with regions containing more than one GGAA motif, nine regions overlapped with areas reported to contain a single GGAA instance.

Binding and geometry frequencies

The analysis of the frequencies of binding site geometries and of ChIP seq binding aims at identifying putative transcription factor modules. Technically, the frequencies measured in both cases can be interpreted as conditional probabilities, namely

$$P\left(B_x \middle| D_{ETS,E2F}\right) \tag{1}$$

the probability of observing a binding event B_x (where x denotes binding of either E2F3 or EWSR1/FLI1) conditioned on the presence of a motif pair ETS and E2F at distance $D_{ETS,E2F}$ and

$$P\left(D_{ETS,E2F} \middle| B_x\right) \tag{2}$$

the probability of the presence of a motif pair ETS and E2F at distance D conditioned

of a binding event overlapping the pair. Intuitively, one might be led to think that these observables are not independent and, in fact, measure the same phenomenon. A closer analysis reveals, that this is not the case: even though the observables (1) and (2) are related by Bayes theorem

$$P\left(B_x|D_{ETS,E2F}\right) = \frac{P\left(D_{ETS,E2F}|B_x\right)P\left(B_x\right)}{P\left(D_{ETS,E2F}\right)} \quad (3)$$

where the structure of the denominator makes that expression non-trivial. The genome wide probability for binding, $P(B_x)$, is a constant for E2F3 and EWSR1/FLI1, respectively. The expression in the denominator is the genome wide probability of observing a tandem ETS/E2F motif at distance D, regardless of a binding event. Empirically, this factor has a complex structure (**Fig S3b**), which in part recapitulates the form of the conditional $P(D_{ETS,E2F} \mid B_x)$. Essentially, equation (3) states that the increased binding probability on the left side of the equation is not observed because, but despite the fact that specific ETS and E2F motif configurations are over-represented in the genome.

Custom software was used to estimate the conditional probabilities (1) and (2). Transcription factor binding sites were identified based on the matrix definition and cut-offs (minFP-Vertebrate) V$CETS1P54_03 and V$E2F_Q2 from the TRANSFAC ver. 11.4 database. These matrices were chosen as representative examples based on their relatively high frequency in the genome, compared to other ETS and E2F matrix definitions. For (1), each binding region was analyzed for the presence of V$CETS1P54_03. If this motif was found, DNA 2000bp up- and down-stream from that match (where downstream was oriented to coincide with the 3'-5' direction of the ETS match) were analyzed for the presence of V$E2F_Q2. Whenever an instance of the E2F motif was identified at distance D from the ETS motif, a counter for the bin at D was increased. For (2), the counting procedure started by identifying all occurrences of V$CETS1P54_03 in the genome, testing if this match was overlapped by a binding site, and increasing a counter for each bin representing E2F motifs at distance D. In order to reduce stochastic noise, a +/- 12 running average low-pass kernel was used on the resulting histograms.

Supplemental Figure Legends

Suppl. Figure SF1: (a) Average relative tag density for input (non-selected) DNA close to transcription start sites. The density of read tags for non-selected input DNA close to transcription start sites. The curve shows a bias towards increased read densities close to transcription start sites (and, most likely, other regions) in the genome.

(b) Representative binding regions.

Tag densities of EWSR1/FLI1 and E2F3 are shown in genomic regions adjacent to known ETS targets *CAV1, GLI1, IGFBP3, STYXL1, NR0B1* and *TGFβRII*.

(c,d) Clustering of FLI1 binding: (c) The distribution of the number of genes as a function of the number of binding events for E2F3 (blue) and EWSR1/FLI1 (red) differs substantially between the two transcription factors. Of the target genes, 37% (3,102) had more than one associated EWSR1/FLI1 localization site, and more than 10% (970) had 4 or more adjacent binding regions. In comparison, the distribution of E2F3 on A673 chromatin was more flat with at most 5 binding sites adjacent to any single gene. Only 10% (398) of genes had more than one, and less than 0.5% (12) of genes had more than 3 E2F3 binding regions.

(d) In this example of a FLI1 binding cluster, the 960kb region inside the gene DLGAP1 contains 15 discrete EWSR1/FLI1 binding regions. The expected number of binding events in a regions o this size is less than one.

Suppl. Figure SF2: (a) Families of transcription factors: Hierarchical clustering of transcription factor motifs identified as enriched in EWSR1/FLI1 regions with at least 50 occurrences. **(b) Genomic background of the ETS/E2F geometry is not trivial:** A genome wide analysis of the frequency of ETS/E2F motif configurations in the human genome, regardless of binding of EWSR1/FLI1 and/or E2F3.

Suppl. Figure SF3: Validation of EWSR1-FLI1 and E2F3 binding to two selected target genes in two additional EWSR1/FLI1 positive ES cell lines. ChIP assays were performed in TC71 and TC252 ESFT cells. EWSR1/FLI1 and E2F3 ChIP were followed by qPCR amplification of promoter regions containing ETS and/or E2F sites,

and, for negative control, a region upstream of the corresponding TSS not containing E2F or ETS binding sites for (**a**) *ATAD2*, (**b**) *E2F3,* and (**c**) GEMIN4. For control, ChIP using unrelated IgG was performed.

Suppl. Figure SF4: Validation of EWSR1-FLI1 and E2F3 binding to three selected target genes. ChIP assays were performed in A673 ESFT cells. Cells were either left untreated (+) for control, or were treated for 48h with doxycycline to induce knockdown of EWSR1-FLI1 (-). (**a-c**) EWSR1/FLI1 ChIP and (**d-f**) E2F3 ChIP were followed by PCR amplification of promoter regions containing ETS and/or E2F sites, and, for negative control, a region upstream of the corresponding TSS not containing E2F or ETS binding sites for (**a, d**) *E2F3*, (**b, e**) *ATAD2* and (**c, f**) *GEMIN4*. For control, ChIP using unrelated IgG was performed. **(g,h)** Fold changes in reporter activity of wild type and mutant reporter constructs for *E2F3* and *GEMIN4* in the presence (+) and doxycycline-induced absence (-) of EWSR1/FLI1 48h after EWSR1/FLI1 shRNA induction

Suppl. Figure SF5: Example of a highly over-represented KEGG pathway "Axon Guidance". Genes with adjacent EWSR1/FLI1 binding signals (highlighted in red) are significantly enriched ($p < 0.001$) in this pathway compared to a flat background model.

Suppl. Figure SF6: Transcription factor motif analysis in **(a**) RPWE+ERG cells and **(b)** VCaP shows significant enrichment for E2F recognition sites in ERG binding regions **(c)** Preferred geometries of ERG binding regions in RPWE+ERG cells measuring location of E2F recognition sites relative to predicted ETS motifs shows a pattern nearly identical to that obtained for EWSR1/FLI1 binding regions in Ewing's sarcoma cells.

Suppl. Figure SF7: (a) E2F3 occupancy of the promoter *ATAD2* is significantly reduced in A673 with a mutated ETS recognition site in comparison to wild type sequence. (b) Mutation of the E2F binding site does not reduce EWSR1/FLI1 binding to the *ATAD2* and *GEMIN4* promoter.

Supplemental Tables

Suppl. Table ST1: Relative increase in E2F expression in primary ES versus mesenchymal stem cells, and in MSC upon ectopic EWSR1/FLI1expression. Shown are log2-fold changes derived from Affymetrix arrays.

Suppl. Table ST2: Statistics for sequence tags obtained in the E2F3, EWSR1/FLI-1 ChIP and non-selected (input) sequencing experiments in A673 and VCaP cells. The columns x/INPUT calculate the ratio of hits observed in the ChIP experiment over the number of tags expected from the input DNA. Additional sheets show per sequencer lane statistics for the various antibodies.

Suppl. Table ST3: Binding regions identified for EWSR1/FLI1 in A673 cells, and E2F3 in A673 and VCaP cells (one per sheet). In addition to genomic coordinates, a characterization of the binding event is included: gene identifiers associated with the binding event (Entrez Gene-ID, Gene Symbol), the distance from the outer boundaries of that gene (Distance), the average tag-count (Amplitude) in the binding region, the peak tag count for both ChIP (Amplitude,maxAmplitude) and input (bgAmplitude, maxBgAmplitude) as well as average phast conservation score (Conservation) and maximum conservation score (MaxConservation).

Suppl. Table ST4: Gene symbols for genes associated with the three dominant PCA components with $|r| > 0.8$ (sheets 1-6).

Suppl. Table ST5: Gene ontology terms significantly over-represented (sheets 1-6) in the gene sets of Supplemental Table ST4.

Suppl. Table ST6: Over-represented binding motifs for EWSR1/FLI1 and E2F3 binding in A673 and for E2F3 in VCaP (sheets 1-3). In this table, background frequencies were estimated for the entire human genome, foreground provides the number of instances of finding the motif inside the respective binding regions. The column “over” provides the estimated over-representation of the motif, and “p” is the associated p-value for that over-representation.

Suppl. Table ST7: Over-represented GO terms for genes adjacent to binding events for E2F3, EWSR1/FLI1 and ERG independently, E2F3 with the respective ETS factor simultaneously, and EWSR1/FLI-1 without simultaneous E2F3 binding in A673 and VCaP cells (sheets 1-6).

Suppl. Table ST8: ETS/E2F3 binding core: regions bound by both E2F3 and the respective ETS factor shared in A673 and VCaP cells.

Suppl. Table ST9: Genes positively regulated by the relevant ETS fusion gene in both A673 and VCaP (left column) and the subset displaying simultaneous ETS and E2F3 binding.

Supplemental References

Abaan OD, Levenson A, Khan O, Furth PA, Uren A, Toretsky JA. 2005. PTPL1 is a direct transcriptional target of EWS-FLI1 and modulates Ewing's Sarcoma tumorigenesis. *Oncogene* **24**(16): 2715-2722.

Beauchamp E, Bulut G, Abaan O, Chen K, Merchant A, Matsui W, Endo Y, Rubin JS, Toretsky J, Uren A. 2009. GLI1 is a direct transcriptional target of EWS-FLI1 oncoprotein. *J Biol Chem* **284**(14): 9074-9082.

Fuchs B, Inwards CY, Janknecht R. 2004. Vascular Endothelial Growth Factor Expression is Up-Regulated by EWS-ETS Oncoproteins and Sp1 and May Represent an Independent Predictor of Survival in Ewing's Sarcoma. *ClinCancer Res* **10**(4): 1344-1353.

Fukuma M, Okita H, Hata J, Umezawa A. 2003. Upregulation of Id2, an oncogenic helix-loop-helix protein, is mediated by the chimeric EWS/ets protein in Ewing sarcoma. *Oncogene* **22**(1): 1-9.

Gangwal K, Lessnick SL. 2008. Microsatellites are EWS/FLI response elements: genomic "junk" is EWS/FLI's treasure. *Cell Cycle* **7**(20): 3127-3132.

Guillon N, Tirode F, Boeva V, Zynovyev A, Barillot E, Delattre O. 2009. The oncogenic EWS-FLI1 protein binds in vivo GGAA microsatellite sequences with potential transcriptional activation function. *PLoS ONE* **4**(3): e4932.

Hahm KB, Cho K, Lee C, Im YH, Chang J, Choi SG, Sorensen PH, Thiele CJ, Kim SJ. 1999. Repression of the gene encoding the TGF-beta type II receptor is a major target of the EWS-FLI1 oncoprotein. *NatGenet* **23**(2): 222-227.

Kikuchi R, Murakami M, Sobue S, Iwasaki T, Hagiwara K, Takagi A, Kojima T, Asano H, Suzuki M, Banno Y et al. 2007. Ewing's sarcoma fusion protein, EWS/Fli-1 and Fli-1 protein induce PLD2 but not PLD1 gene expression by binding to an ETS domain of 5' promoter. *Oncogene* **26**(12): 1802-1810.

Kinsey M, Smith R, Lessnick SL. 2006. NR0B1 is required for the oncogenic phenotype mediated by EWS/FLI in Ewing's sarcoma. *Mol Cancer Res* **4**(11): 851-859.

Luo W, Gangwal K, Sankar S, Boucher KM, Thomas D, Lessnick SL. 2009. GSTM4 is a microsatellite-containing EWS/FLI target involved in Ewing's sarcoma oncogenesis and therapeutic resistance. *Oncogene* **28**(46): 4126-4132.

Nakatani F, Tanaka K, Sakimura R, Matsumoto Y, Matsunobu T, Li X, Hanada M, Okada T, Iwamoto Y. 2003. Identification of p21WAF1/CIP1 as a direct target of EWS-Fli1 oncogenic fusion protein. *JBiolChem* **278**(17): 15105-15115.

Prieur A, Tirode F, Cohen P, Delattre O. 2004. EWS/FLI-1 silencing and gene profiling of Ewing cells reveal downstream oncogenic pathways and a crucial role for repression of insulin-like growth factor binding protein 3. *MolCell Biol* **24**(16): 7275-7283.

Siligan C, Ban J, Bachmaier R, Spahn L, Kreppel M, Schaefer KL, Poremba C, Aryee DN, Kovar H. 2005. EWS-FLI1 target genes recovered from Ewing's sarcoma chromatin. *Oncogene* **24**(15): 2512-2524.

Smith R, Owen LA, Trem DJ, Wong JS, Whangbo JS, Golub TR, Lessnick SL. 2006. Expression profiling of EWS/FLI identifies NKX2.2 as a critical target gene in Ewing's sarcoma. *Cancer Cell* **9**(5): 405-416.

Tirado OM, Mateo-Lozano S, Villar J, Dettin LE, Llort A, Gallego S, Ban J, Kovar H, Notario V. 2006. Caveolin-1 (CAV1) is a target of EWS/FLI-1 and a key determinant of the oncogenic phenotype and tumorigenicity of Ewing's sarcoma cells. *Cancer Res* **66**(20): 9937-9947.