Figure S1: Improper mapping of a peak location from ChIP-chip data. The ChIP-chip experiment (top panel, array identifier number 82998_GM06990_e2f4b_ratios.gff from Xu et al. Genome Res. 17:1550-61, 2007) clearly identified the GMNN promoter as an E2F4 target and follow-up ChIP-PCR experiments (using relatively large chromatin fragments) confirmed binding of E2F4 to the GMNN promoter region. However, when the region corresponding to the peak identified by ChIP-chip was cloned into a plasmid and analyzed as an isolated region, very low binding of E2F4 was detected (A Rabinovich and PJ Farnham, unpublished data). Subsequent analysis using ChIP-seq revealed that the center of the peak (and by implication the sequences corresponding to the binding site) is offset from the region represented on the promoter microarray. Thus, while both technologies correctly identify the GMNN promoter as a target for E2F4, ChIP-seq provided a more accurate location of the binding site. As is typical for the different technologies, the enrichment values plotted on the Y axis for ChIP-chip data are on a log2 scale whereas the tag counts plotted on the Y axis for the ChIP-seq data are on a linear scale. The region of chromosome 6 containing the GMNN gene is shown; the peak is over the 5' end of the gene.

Figure S2: ChIP-seq mapping protocols can result in loss of certain peaks.
The promoter region of the FAM72B gene was identified as an E2F4 target using ChIP-chip (**panel A**). However, there was no binding site visible in the ChIP-seq data (**panel B**). Insight into why this peak was missed by ChIP-seq comes from an analysis of the uniqueness of each 27mer in the shown region. Most 27mers in the FAM72B promoter region are present in 3 or more copies in the human genome (**panel C**). Although in certain cases, sequencing longer reads may rescue lost peaks due to non-uniqueness of a particular tag, this approach would not rescue the FAM72B peak. As shown in the **inset**, only 37% of the tags in the peak region would be unique if 60mers were analyzed. Another approach to rescue the missing peak is to allow tags that match to less than 10 places (with 0 or 1 mismatch) in the genome to be included in the analysis. If this is done, the FAM72B promoter peak is recovered (**panel D**). However, under these conditions, each non-unique tag would be mapped to all matching locations. This allows a peak to be identified, although one could not know which of the chromosomal locations was in fact bound by the factor. Follow-up ChIP-PCR using primers that are unique to one of the regions may be possible, assuming

that some slight sequence divergence has occurred. Another problem is that thousands more peaks were identified in the ChIP-seq dataset when tags mapping less than 10 times were included in the binning. Clearly, it is not favorable to obtain many false positives when recovering a few false negatives. However, if the number of times a tag is found in the genome is taken into account, the promoter peak remains but many of the false positives are removed (**panel E**).

Figure S3: ChIP-seq experimental protocols result in non-uniform genomic analyses. An important component of the precise peak definition that is obtained using ChIP-seq is that the ChIP sample is cut out of a gel, providing a sample that is composed only of fragments of a particular size range (usually 150-400 nt in length). However, certain regions of human chromosomes are not easily fragmented to this small size; this is especially problematic for heterochromatic regions. Therefore, if a peak is in such a region, the chromatin fragments enriched in the ChIP experiment will be longer than 400 nt and subsequently discarded in the gel isolation step, resulting in under-representation in the sequencing library. This problem can be illustrated using a library prepared from an input chromatin sample. Two different size library fragments from a K562 input library, 150 to 400 bp and 400 bp to1 kb, were extracted from a gel (**panel A**) and analyzed by PCR (**panel B**). Regions that that are located within heterochromatin (such as the indicated ZNFs) were represented more in the larger size fragment, while targets located within euchromatin (such as LSM4 and PPWD1) were equally present in both size fragments. Because the density of tags in the genomic regions containing the indicated ZNF genes is less than would be expected, peak identification will be adversely affected unless a proper analysis is performed. Since input libraries show the same bias as ChIP libraries, binding sites should be determined as regions that are significant over background, independent of sequence density.

Figure S4: False positive peaks can occur upon analysis of ChIP-seq data. Shown is the pattern of binding of two unrelated site-specific transcription factors and a track of input DNA from K562 cells. The peaks enclosed by the red and green circles are found in both ChIP-seq datasets and in the input library. The peak outlined by the green circle (see close up in right bottom panel) corresponds to pericentromeric DNA and the peak outlined by red

circle (see close up in left bottom panel) corresponds to a false positive that is most likely due to a small region that is not unique in the human genome. The Sole-Search peak-calling program elminates both of these types of false positives.

Figure S5: Input and Output files for Sole-Search, the Gff-Overlap Tool, and the Location-Analysis Tool. Shown are the files that are uploaded into each of the programs and a list of the output files produced by each program. The output files from Sole-search, the Overlap-Analysis Tool, and the Location-Analysis Tool for the various ChIP-seq datasets are in the **Supplementary Data Folders (**see **List of Supplementary Materials** for a description of the contents of each **Supplementary Data Folder).**

Figure S6: E2F4 output files from the Location-Analysis Tool. The significant peaks file obtained after merging two ChIP-seq replicate data sets for E2F4 in K562 cells was uploaded into the Location-Analysis Tool and the information contained within the resultant output files was graphed. Shown are the number of binding sites per chromosome (**A**), the number of binding sites per gene (**B**), the distance of the set of binding sites from the start site of transcription (**C**), the location analysis (**D**) with respect to the start and stop codon of a gene (the different designations are gd: gene desert (gd) includes regions greater than 100 kb from a refseq gene, 5d includes regions between 10 kb and 100 kb upstream of a refseq gene, 5p2 includes regions between 2 kb and 10 kb upstream of a refseq gene, 5p1 includes regions that are less than 2 kb upstream of a refseq gene, gene includes any exon or intron, 3p1 includes regions less than 2 kb downstream of the last exon of a refseq gene, 3p2 includes regions between 2 kb and 10 kb downstream of the last exon of a refseq gene, 3d includes regions between 10 kb and 100 kb downstream of the last exon of a refseq gene), and the distribution of the intragenic binding sites (**inset in D**).

Figure S7: TCF4 output files from the Location-Analysis Tool. The significant peaks file obtained after merging two ChIP-seq replicate data sets for TCF4 in HCT116 cells was uploaded into the Location-Analysis Tool and the information contained within the resultant output files was graphed. Shown are the number of binding sites per chromosome (**A**), the number of binding sites per gene (**B**), the distance of the set of binding sites from the start

site of transcription (**C**), the location analysis (**D**) with respect to the start and stop codon of a gene (the different designations are gd: gene desert (gd) includes regions greater than 100 kb from a refseq gene, 5d includes regions between 10 kb and 100 kb upstream of a refseq gene, 5p2 includes regions between 2 kb and 10 kb upstream of a refseq gene, 5p1 includes regions that are less than 2 kb upstream of a refseq gene, gene includes any exon or intron, 3p1 includes regions less than 2 kb downstream of the last exon of a refseq gene, 3p2 includes regions between 2 kb and 10 kb downstream of the last exon of a refseq gene, 3d includes regions between 10 kb and 100 kb downstream of the last exon of a refseq gene), and the distribution of the intragenic binding sites (**inset in D**).

Figure S8. Location analysis of E2F4 peaks identified using different programs
The signifpeaks.gff files for the merged E2F4 datasets called by Sole-search (A), PeakSeq (B), Sissrs-using background correction (C), and Sissrs –with no background correction were uploaded into the Gff-Overlap Tool, using 0 nt distance. A graphical representation of the dist_analysis file obtained by analyses of each set of peaks is shown.

Figure S9. Illustration of the consequences of removal of individual steps of Sole-search. The Sole-Search software includes three steps, which, when combined, produce statistically significant peaks. Removal of any individual step in the process increases the number of false positive peaks called. Removal of the first step (panel A), which normalizes the genome based on copy number, increases the number of false positive peaks called in duplicated regions of the genome. Removal of the second step (panel B), which determines the statistically significant height cutoff, increases the false discovery rate dramatically by including many small “peaks” that are enriched over a background. Removing the third step (panel C), which determines significance over background, increases the number of small, presumably false positive, peaks called and does not narrow the binding sites.