|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Cell Line | Cancer | TF ChIP-seq | Histone ChIP-seq | EnhancerSeq | RNA-seq |
| K562 | CML | 209 | 12 | Y | Y |
| GM12878 | Normal Lymphoblastoid | 101 | 11 | Y | Y |
| HepG2 | Liver Cancer | 97 | 11 | N | Y |
| MCF-7 | Breast Cancer | 51 | 5 | Y | Y |
| A549 | Lung Cancer | 32 | 11 | N | Y |
| HeLa-S3 | Cervical Cancer | 60 | 11 | N | Y |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Cell Line | eCLIP | DNase-seq | Hi-C | MMBS/RRBS | RepliSeq | siRNA/shRNA RNAseq |
| K562 | Y | Y | Y | Y | Y | Y |
| GM12878 | N | Y | Y | Y | Y | N |
| HepG2 | Y | Y | N | Y | Y | Y |
| MCF-7 | N | Y | Y | Y\* | Y | Y |
| A549 | N | Y | N | Y\* | N | N |
| HeLa-S3 | N | Y | N | Y\* | Y | N |

Using the ENCODE regulatory data to interpret non-coding somatic variants

Authors in this paper include members from several labs from within ENCODE, including the Gerstein lab, the White Lab, the Liu Lab, the Klein Lab, the Yue lab, and the Gilbert Lab

We understand the impact of somatic mutations well in a only very limited number of cancer genes; in contrast, the overwhelming number of mutations in cancer genomes occur in non-coding region. The new release of the ENCODE data allow us to bridge these two facts. First, the new ENCODE data enables precise tissue-matched genome-wide background mutation rate calibration in a variety of tumors by separating the effect of well-known confounders, such as replication timing and chromatin status. Furthermore, by integrating large scale of ChIP-seq, DNase-seq, Enhancer-seq, Hi-C and ChIA-PET data from ENCODE, we are able to define with high confidence distal and proximal regulatory elements and their linkages to annotated genes. This enables us to create extended gene definitions, and we are able to show these are more sensitive than coding region only analysis in terms of burdening. In particular, in leukemia it allows us to find well-known drivers such as TP53 and ATM, but also pick up other key genes such as BCL6 [[pick a different cancer – todisc later?]], which can then be associated with patient prognosis. Second, we integrated the ENCODE data to build up a high confidence TF-gene regulatory network. This enabled us to identify highly rewired (i.e. target changing) TFs, such as NRF1 and MYC in comparing tumor and normal samples. By integrating large-scale chromatin features, we demonstrated that such massive rewiring events between tumor and normal cell lines are mainly attrributable to the chromatin structure changes instead of direct mutational effect. Furthermore, we also found that TFs with more mutationally burdened binding sites (eg [[fill in TF names]])) tend to be located at the bottom hierarchy of the TF regulation network. Third, using the ENCODE regulatory network, we developed integrative scoring workflow to prioritize key elements (and mutations in them) according to their role in cancer and then validated these in small-scale studies. [[now unified this sect. ]] In particular, we prioritized ZNF687 as a key TF for breast cancer and SUB1 as a key RNA binding protein for liver and lung cancer and validated them through siRNA knockdown experiments. Finally, we identified key enhancers and mutations in them in breast cancer and then validated the functional effect of the mutations through luciferase assays.