Requirements for a DNase-seq and FAIRE-seq experiments

Following an analysis of deeply sequenced DNase-seq and FAIRE-seq datasets, we suggest the following requirements for a DNase-seq and FAIRE-seq experiments.

1. Limited number of sequenced reference samples
   In some cases, deeply sequenced reference samples such as Input DNA exhibit "peaks" in the promoter regions, as well as other regions of the genome that is under-represented in the reference sequence (alpha satellite regions, ribosomal subunits, etc.). We believe that the input promoter peaks are a result of either endogenous nuclease activity, or FAIRE-like regions that are not cross-linked during the sample preparation. These promoter peaks likely represent open chromatin and should not be excluded from our analysis. The DNase/FAIRE-seq data itself can be used to identify regions that exhibit copy number variations in our samples or that are under-represented in the sequenced reference genome. The true signals from FAIRE and DNase exhibit a unique sharp signal structure that differs radically from the type of signal produced by these genomic features. Therefore, for DNase and FAIRE experiments, we will not require input sequencing from additional cell types.

2. Depth of Sequencing
   Since DNase and FAIRE sites represent a continuum of chromatin openness, achieving true "saturation" may not be practical. However, a decision must be made regarding adequate level of coverage. We propose that the optimal depth of sequencing be guided by our ability to identify regions that were also identified other methods such as by tiled arrays (ENCODE pilot arrays or equivalent), qPCR, or Southern blots.

3. Number and Reproducibility of Biological Replicas
   By definition at least two biological replicates are necessary to ensure that the experiment is reproducible. However, experiments completed to date indicate that there will not be a significant gain in information beyond two biological replicates, when they are in reasonable agreement. We propose to require the following agreement between biological replicas:
   a) The number of mapped reads from replicas should be within a factor of two of each other.
   b) The length of target lists should be within a factor of two of each other.
   c) Either of the following options
      I. Intersect top fraction (40%) of target list from one replica against the entire other target list and require a threshold amount of overlap (90%). Repeat for the reciprocal.
      II. Target lists scored using all available reads from both replicates must share more than 75% of targets in common with each of the replica experiments.

These parameters will have to be revisited as more data sets are available, and methods to compare replicates of different quality or enrichment levels are developed..