PI: Stamatoyannopoulos, J

Project Title: **A comprehensive catalogue of human DNaseI hypersensitive sites**

I. Lists of data types

I.1) DNase-array data. This is generated during secondary Q/C and comprises a single Affy whole-genome tiling 2.0 chip (one of the 7 in the set, varying by cell type).

I.2) Digital DNaseI data. This is generated on the Illumina/Solexa platform (and soon on the ABI platform). A large amount and variety of data files are generated and can be supplied either in their entirety or in a subset to the DCC. It is expected that initially uniquely mapping reads will be supplied to the UCSC browser for display.

I.3) DNaseI hypersensitive sites. These are locations of DNaseI hypersensitive sites (chr/start/stop) as determined from digital DNaseI data. These site coordinates are used in the assay performance characteristics computed independently using the validation data.

I.4) DNaseI Hypersensitivity Southern data. These are digital images from Molecular Dynamics Phosphorimager scans. A fixed set of 424 probes regions (HindIII fragments) distributed in stratified random fashion around the genome is used. The location of these regions will be available on our website and presumably in information fields with the appropriate repository has been set up. We deposit the sequences of probes themselves in UniSTS.

II. Planned number of biological replicates for verification of primary data

For each cell type studied, 16 biological replicates will be generated. All of these will be subjected to primary Q/C to assess background digestion with real-time PCR. ~14 replicates per cell type will advance to secondary Q/C to be screened for average signal-to-noise ratio using tiling DNA microarrays. Of these, the top 3 replicates passing a stringent minimum quality threshold will be selected for deep sequencing. We anticipate that some cell types may only have 2 samples passing the high secondary Q/C threshold.

III. Plans for release of verified data

DNase-array data are scheduled for release to GEO immediately after generation (in practice this may take 2-3 weeks until they are finally uploaded). These data do not need to be represented in the UCSC browser since they are Q/C assays.

Digital DNaseI data will be released following verification. A verified data set is defined as two or more biological replicates that have been sequenced to saturation (>25million uniquely mapping reads), and which display average correlation coefficient between their density

functions (computed in 150bp window sliding q20 bases) of >0.94. When two samples diverge by more than this, additional samples will be obtained and sequenced.

IV. Plans for release of validated data - six month maximum interval after verification was proposed

Validation is independent of the above and is not a gating item for data release. Validation data = the sensitivity/specificity/PPV/NPV of digital DNaseI applied to each cell type will be computed *a posteriori* when a sufficient number of Southern blots have become available to make these computations with reasonable power. We anticipate that interim analyses will be possible approximately every 4 months and will cover multiple cell types. Southern validation data will be released on a continuous basis and can commence as soon as a suitable data repository and visualization tool is available.