We list some possible features in rough order of perceived utility.

DNase

The DNase data should predict the regions with binding activity reasonably well. Correlations between maximal DNase levels and binding range roughly from .03 to .4.

Region level summaries

- Maximal value in region
- Average value in region

Higher resolution

We have DNase data down to base-pair resolution that could be used although the length-scale of the data seems to be somewhere inbetween base-pair and region size (200bp).

DNase footprinting

Technique to narrow down DNase reads to tight footprint of TF binding site. See Wellington, a tool by Sasha's group.

Sequence

Motifs

We can scan for motifs and summarise the hits for each motif in each region.

- Maximum score
- Average score
- Number of hits
- Any other combination of hits, such as that used in `BiFa`

It could be worthwhile to combine/reweight the motif hits using the DNase footprinting data.

Known motifs

- JASPAR
- CisDB
- HOCOMOCO
- Jolma et al. (2013)
- . . .

de novo motifs

What positive/negative sequences to use? For the negative sequences we could use the ambiguous regions with highest DNase scores as these will be hardest to discriminate.

DREME discovers very short motifs (`W=3`) for some TFs that do not have hits above the `steme-pwm-scan` genome-wide threshold.

Spacing between sites

Once we have motif sites we can look for significantly enriched spacing between them.

Sequence shape

Can be calculated by `R` package. Is supposed to be predictive of binding above and beyond motifs. Could a neural network compensate for this with sequence level data?

Expression

The expression data should be complementary to the other sources of data. It is not clear that it will be useful without gene annotations. Dimensionality reduction could be used to extract features.

Annotations

Using expression data with annotations we can define for each region

- Expression levels of nearest gene
- Average expression levels of nearby genes
- Expression levels of TFs predicted to bind motif hits in the region