1 THESIS PLAN

1.1 INTRODUCTION

Need to introduce the seminal studies of genome organisation, particularly those using C methods (Lieberman-Aiden, Dixon, older 3C stuff etc.). Detail the fractal globule model of genome organisation, as well as counter theories like strings, binders and switches (SBS). There's also growing computational literature in calling domains, sub-domains (e.g. Bing Ren's D.I. + HMM, but also alternating domains, breakpoint algorithm etc.).

Also mention the criticisms of C methods — see references mention by Wendy (PLOS ONE) — and caveats of processing the data (biases, normalisation etc., start with Yaffe and Tanay biases, ICE, others).

Machine learning has been successful at calling chromatin states (`ChromHMM`, `SegWay`, others) and generally building models of complex biological phenomena. Explain how this type of computational approach has led to biological insights. with particular deference to recent ENCODE studies such as Dong *et al.* (2012).

1.2 METHODS

I already have some text for this from first year report and the paper methods section. Things to cover include:

- Processing raw reads, mapping
- ICE, Hi-C normalisation
- Calling boundaries, HMMs
- Modelling, random forests, variable importance
- GLASSO, regularisation
- Citations for R packages used
- Package code for entire thesis?

1.3 RESULTS

1.3.1 **Early stuff:** Modelling transcription and chromatin

I replicated the work of **(author?)** (1) in modelling of transcriptional output based on a large set of ENCODE features. I extended their work by adding new features, and dissected the "best bin" approach to discover where (relative to a gene) influential variables correlated best with expression.

We then applied the same approach to modelling a different set of data: the A / B compartment profiles reported in Lieberman-Aiden *et al.* (2009). Noting that Hi-C datasets were available for the three tier 1 ENCODE cell lines, I applied this modelling approach to each in turn, with their own corpus of ENCODE features.

1.3.2 **Model dissection**: regularised models, influential variables, cross-application

Having reasonably accurate models of chromatin organisation, it's then of interest to understand why they are successful and if improvements can be made. Firstly, rankings of variable importance were looked at per cell type model. Models were also cross-applied form one cell type to another.

We were interested in building minimal viable models, or those suitably regularised such that accuracy was maintained while the dimensionality of input features was minimised. To this end, we employed the Graphical LASSO algorithm, a tuneable L1 regulariser, to reduce each model of 36 variables down to approximately 5 with little loss in predictive power. However, this "wrapper" method of regularisation was independent from the learning algorithm, hence may not represent a truly optimal subset of features.

In order to resolve this, we employed a regularised Random Forest algorithm, as well as a brute-force process of constructing all possible subset models with varying numbers of features. For example, all combinations of five variables from the original 36 were passed to the learner and the accuracy was compared. From this we discovered that while model performance was affected by the number of input features, due to the pervasive multicollinearity any subset model of five variables would perform almost equally well. This signalled thta generated minimal viable models may provide little additional understanding of the relationships between higher order chromatin organisation and our locus level features.

1.3.3 **Odds 'n' ends**: TADs, boundaries, super bounds, G-bands

TADs are a well-described facet of higher order chromatin organisation at a scale below that of nuclear compartments. We recalled these domains in each cell type under study and compared the results. Unsurprisingly we found TAD boundaries to be well-matched between these cell types, confirming them as a relatively invariant level of organisation.

The boundaries of TADs have previously been reported as bound by numerous factors, some of which (e.g. CTCF) have previously implicated roles in organising genome conformation. With a larger set of ChIP-seq datasets available, we quantitatively tested for enrichment or depletion of 36 DNA binding proteins and histone modifications. This enabled us to compare enrichments across cell types and identify those that were consistently marking these boundaries. Further, we applied the same methodology to boundaries of compartments and discovered similar spectra of enrichments and depletions, but at a lower resolution — in agreement with a "fractal globule" view of genome organisation.

We investigated the idea of "*super boundaries*" which were both TAD and compartment boundaries. It emerged that these boundaries, though present, did not display stark differences from non-overlapping TAD or compartment boundaries.

We also found an agreement between A and B compartments with the long-known Giemsa stain bands. Both were previously known to correlated with patterns of high and low GC content ("isochores").

1.3.4 Additional chapters from my next project

Now that a paper from my initial project is submitted, I'll be starting a new project which should in theory fill ~2 (?) chapters. This project will continue with the genome organisation theme and likely continue to make use of some of the reprocessed datasets I have generated.

Potential projects include:

- Investigating the contacts between (e.g.) predicted epistatic genes
- Relating Hi-C

1.3.5 Collaborations and side projects

I've also analysed related C-methods data produced by researchers in wet lab groups:

- Adam Douglas (**Hill group**) 4C, Capture-C: Analysis of 4C contacts between the ZRS enhancer and the SHH gene in mouse developing limb bud cells. Treated experiments are awaiting sequencing. A new C-method, Capture-C, will also be used across this region to backup findings from the 4C and FISH experiments.

- Iain Williamson (**Bickmore group**) 5C: Comparing contacts for anterior and posterior developing limb over the HoxD locus. Also comparing with existing mouse Hi-C data to visualise a potentially changing TAD structure.

1.4 DISCUSSION

Summarise key results and place into broader (particularly biological) context.

1.5 END MATERIAL

- Appendix
- References

REFERENCES

[1] Dong X, Greven MC, Kundaje A, Djebali S, Brown JB, Cheng C, Gingeras TR, Gerstein M, Guigó R *et al.*, 2012. Modeling gene expression using chromatin features in various cellular contexts. *Genome biology*, **13**(9): R53.