Unravelling higher order genome organisation [working title]

Introduction

Benjamin L. Moore

June 11, 2015

igmm
INSTITUTE OF GENETICS
& MOLECULAR MEDICINE

1 REANALYSIS OF HI-C DATASETS

1.1 INTRODUCTION

Since the initial publication of the Hi-C technique in 2009,[?] there has been rapid advancement of both the technique itself and the resolution at which interaction frequencies have been analysed. From the proof-of-concept analysis at 1 megabase (Mb) and 100 killable (kb) resolution,[?] subsequent experiments achieved first 40 kb[?], then 10 kb[?] and most recently 1 kb[?], enabling bona fide fragment-level analysis for the first time.

Such rapid progression in the field has resulted in a wide variety of public Hi-C datasets being available, albeit with differing qualities. With proper correction and at a suitable resolution, these interaction frequencies can be compared and contrasted within and between species.

In this work I uniformly reprocessed publicly-available human Hi-C datasets, in order to address fundamental questions about the stability of higher order genome organisation within cell populations from the same species. Previously Hi-C studies have compared two samples per species, such as K562 against GM06990[?] or IMR90 against GM12878.[?] Here I make use of three Hi-C datasets corresponding to extensively-studied human cell lines: K562, GM12878 and H1 hESC. Together these make up the "Tier 1" cell lines studied by the ENCODE consortium,[?] hence have huge amounts of matched ChIP-seq and histone modification data available.

By combinatorial reanalysis of these cell-matched datasets, I can investigate t

1.2 HI-C REPROCESSING

Each Hi-C dataset used in this work was reprocessed using the same pipeline from raw sequencing reads. In each case, experiments used the same HindIII restriction enzyme.

1.3 COMPARTMENT PROFILES

After uniformly reprocessing each Hi-C dataset and calling compartment eigenvector profiles (see *Methods*), we can compare these between three human cell lines. Compartment profiles have a visibly high-correspondence (Fig. 1), despite the variable sources of both sample material and experimental data.

This close correspondence also validates our approach of combining these different datasets, and suggests our uniform pipeline is successfully accounting for differences in sequencing depth and other batch effects. The precise correlations of these independent measures are in the interval $R = [.75, .8]$ (Fig. **??**; Pearson correlation coefficients, PCC).

1.4 DOMAIN CALLS

The continuous compartment eigenvector can be used as-is to classify A/B compartments, using positive and negative eigenvector values after first orientating the vector with respect to, for example, PolII Chip-seq data.[?]

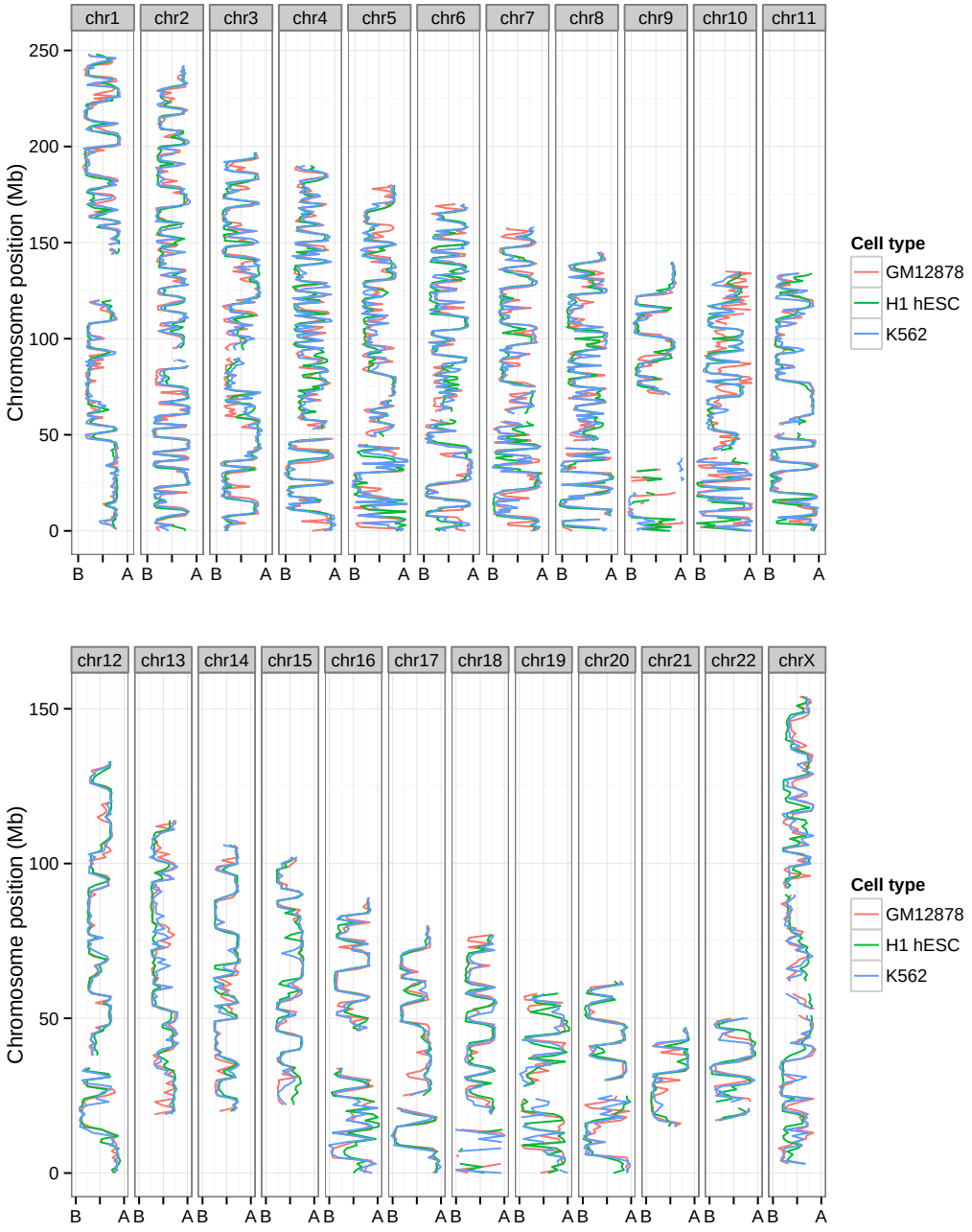

Figure 1: Compartment profiles are observably well-correlated between human cell types and across all chromosomes. Caption

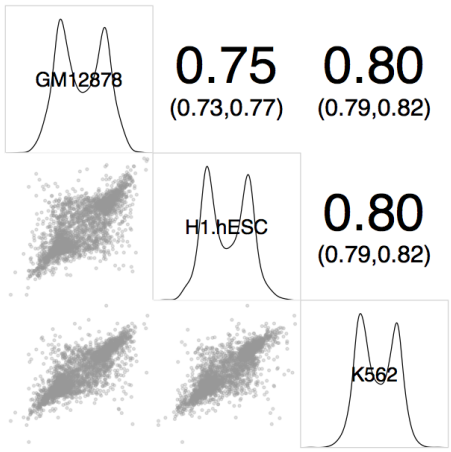

Figure 2: Compartment eigenvectors are well-correlated between human cell types Megabase resolution compartment eigenvector values are shown in a plot matrix. *Upper triangle*: Pearson correlation coefficients between pairs, with 95% confidence intervals (??); *diagonal* Kernel density estimates of eigenvector values per cell type; *lower triangle*: *x*-*y* scatterplot of values.

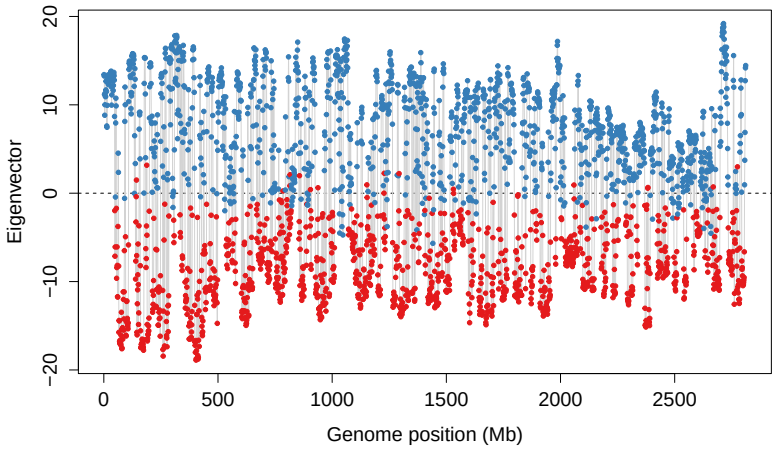

Figure 3: ?? placeholder

However, given the definition of compartments as generally broad and alternating domains along a chromosome, often matching other large domains of Lamin association, an improved classification method might penalise the calls of short compartment calls, which may be the result of noise.

For this reason, instead of using raw eigenvector values we consider observed values as emissions from unobserved underlying states. This can be modelled through a Hidden Markov Model (HMM), whereby we first parameterise models of state and their transitions, then infer the most likely state sequence to have emitted our observed data. This unobserved two-state sequence is then used for compartment calls.

In practice, this acts to de-noise our compartment calls. Where single sign-changes along the series would have resulted in a single-block compartment, these may now be modelled as noisy emissions from a single unobserved state. An examplar region is showing in Fig. 3.

1.5 VARIABLE REGIONS

Despite the vast majority of the genome being in matched chromatin compartments, there are also regions of disagreement. Reasons for observable differences include technical errors and bias, but also more interesting functional explanations, where cell-type specific activation or repression is reflected in changes in higher order structure.

1.6 NUCLEAR POSITIONING