1 LOCAL CHROMATIN CONFORMATION

1.1 INTRODUCTION

The Hi-C assay provides a genome-wide overview of chromatin conformation, however this broad scope imposes resolution limits inherent to an all-vs-all assay. For a closer look at chromatin conformation within a region of interest, alternative C-based assays such as 3C, 4C and 5C can be employed alongside classical microscopy techniques like FISH.

Here I discuss two collaborative projects involving the use of 4C and 5C data to "zoom in" on two well-studied regions related to limb development: the ZRS enhancer and HoxD gene cluster.

1.2 4C AT THE SHH LOCUS

Anterior-posterior patterning in the developing limb is regulated in mammals by *Sonic hedgehog* (SHH).[3] Specifically, the SHH gene is expressed within a confined region named the "zone of polarising activity". Its expression within this region is known to be regulated by a well-studied enhancer, the "zone of polarising activity regulatory sequence" or ZRS.[4] ZRS is located almost 1 Mb downstream of its target SHH promoter in humans, and is located in intronic regions of another gene, LMBR1 (Fig. 1).[4,5] Single point mutations and short insertions within this enhancer have been linked to various limb deformities, including pre- and post-axial polydactyly.[3,5,6] For example, a heritable point mutation in the ZRS enhancer is the cause of polydactyly in "Hemingway cats", a large group of domestic cats with extra toes that reside at the former home of Ernest Hemingway.[6]

Collaborators have developed a model system which allows inducible SHH expression in a non-expressing 14fp cell line derived from the developing limb bud. Application of trichostatin A (TSA) then leads to detectable SHH expression, and increased levels of the histone activation mark H3K27ac at the ZRS (*unpublished data*). However, the question remains whether this TSA treatment is fundamentally altering local chromatin structure, that is, bringing together the ZRS enhancer with its target SHH promoter, or whether ZRS and SHH are in contact in both the active and non-expressing cell lines and SHH expression is blocked through other means. Analysis of the region through FISH implies similar levels of compaction in SHH expressing and non-expressing cells (*data not shown*), suggesting the latter explanation. Addition-

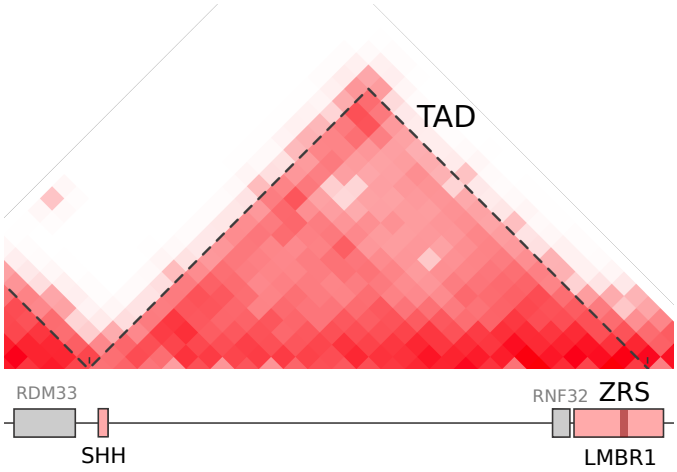

Figure 1: SHH–ZRS contacts occur within a stable TAD. An approximately 1 Mb region of the mouse genome is shown below a Hi-C contact map (derived from previously published data[1]). A clear TAD can be identified spanning from SHH to ZRS, dashed lines show TAD boundaries called by Dixon *et al.*[1]. This figure was generated for Anderson *et al.*[2].

ally, studies have previously reported the SHH–ZRS interaction is an example of a "pre-formed loop", and as such is maintained regardless of transcriptional activity.[7]

My part in this collaboration was to analyse 3C-seq (also known as 4C) data generated by our collaborators for the SHH–ZRS region in mouse. Additionally, the 4C procedure[8] was adapted for specific in-house sequencing instruments (an Ion Torrent Ion Proton™ sequencer as opposed to Illumina™ technology) and as such required diagnostics to confirm the experimental data was accurate.

1.2.1 4C pipeline

The 4C analysis pipeline, starting from de-multiplexed `fastq` files as produced by our in-house sequencing facilities, can be summarised as:

1. Trim known bait sequence using `cutadapt`,[9] select only those reads where known sequence was present
2. Map reads to reference genome `mm9` using `bowtie2`[10] with the `very-sensitive` flag
3. Filter alignments with a MAPQ score < 30 to select for high-confidence alignments using `samtools`[11]
4. Normalise and analyse contacts using the `r3cseq` R package (Methods **??**)

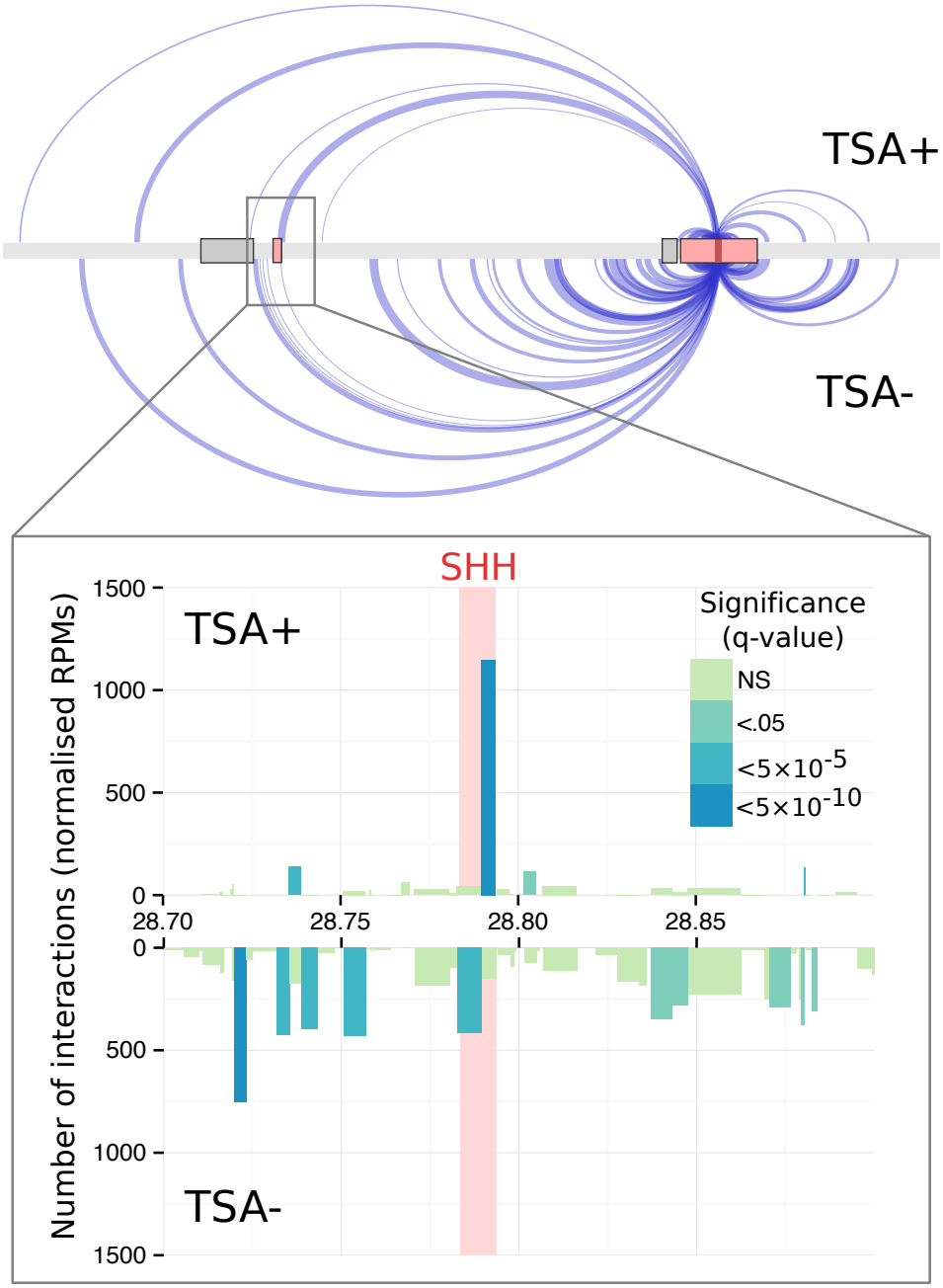

Figure 2: TSA treatment induces a strong ZRS–SHH interaction. 4C interactions are shown as edges from source node (ZRS enhancer bait fragment) to targets along an approximately 2 Mb region of chromosome 5. Edge width is proportional to the number of interactions, only highly significant interactions are shown (FDR q-value $< 5 \times 10^{-5}$). Zoomed region shows the number of interactions of the bait region with SHH in both untreated and TSA treated (after 18h) samples. Each rectangle is a restriction fragment, coloured by FDR q-value indicating the significance of the interaction above expected levels.

1.2.2 Analysis of ZRS interactions

4C experiments were performed by collaborators using the ZRS region as a bait sequence, or "viewpoint", such that it contacts were measured with all other HindIII restriction fragments genome-wide. 4C was performed in both untreated and non-SHH expressing cells (*TSA–*) and in cells treated with TSA, thereby causing SHH expression (*TSA+*).

The first stage in analysing these contacts is to convert observed raw sequencing reads to normalised frequencies (Methods **??**), these normalised values are then assigned significance scores in the form of q-values, with the aim of finding those significantly over-represented relative to expectation (Methods **??**).

The results of a comparison between TSA treated and untreated samples is shown in Figure 2. In it we see a striking and highly significant ZRS–SHH contact in the treated sample (q-value $< 5 \times 10^{-10}$), with a weaker but still significant contact in the adjacent restriction fragment in the untreated sample (q-value $< 5 \times 10^{-5}$).

We also see more broadly a much higher total number of significant contacts in the untreated sample around the viewpoint (Fig. 3). In the treated samples, only a few contacts cross the stringent q-value threshold, with the SHH–ZRS contact among the most significant in terms of both q-value and supporting number of reads (indicated by the thickness of the arc, Figs. 2, 3).

Existing multi-probe FISH data produced by our collaborators shows approximately equal levels of compaction in this region in both TSA treated and untreated cells (*data not shown*). This information in combination with the 4C results reported here (Fig. 2) support a hypothesis that while both samples are held together in a TAD (Fig. 1) — unavoidably inducing many contacts — it is only in the treated sample where a highly-specific ZRS–SHH contact occurs and potentially is then bring about expression of the *SHH* gene.

1.2.3 Assay diagnostics

The 4C protocol used by our collaborators in this work was that of Stadhouders *et al.*[8]. In it, the authors advise some statistical tests to ensure the quality of the experiment results. Among these were:[8]

1. Sequencing reads should be found to have high duplication rates of 95% or greater.

2. 50% or more of all reads should map to the chromosome on which the bait region is located.

Sequence duplication levels were measured with `FastQC`[12] and are shown in Table 1. We find slightly lower than expected levels of duplication, ranging from 62.8%

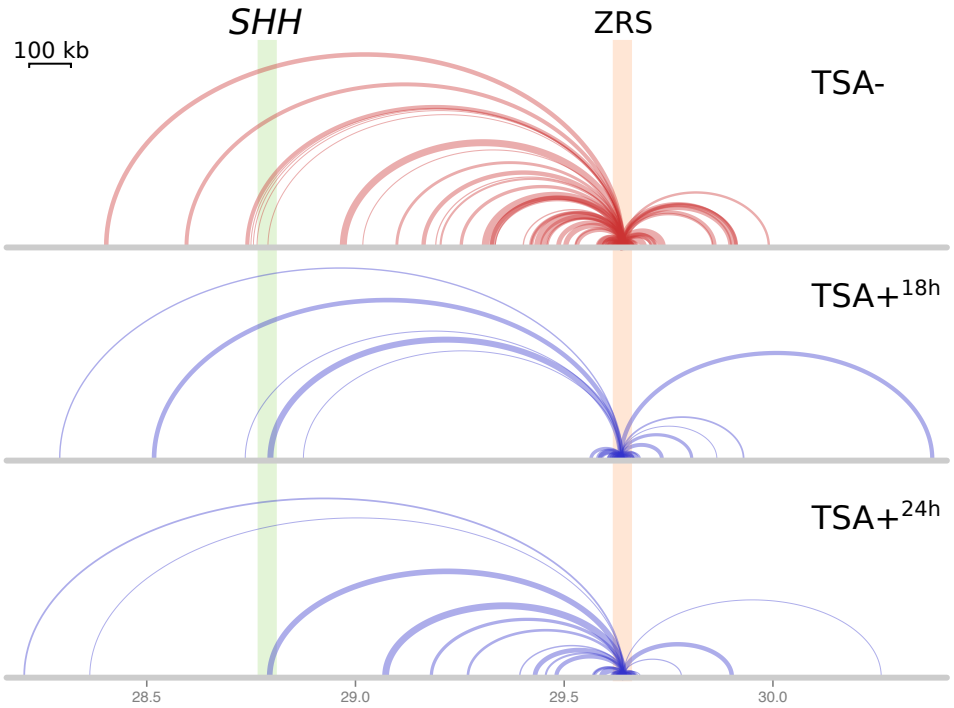

Figure 3: A stable ZRS–SHH interaction is coupled with reduced extraneous contacts. Arc plots are shown for an untreated, non-expressing 14fp cell population (*untreated*) and following TSA treatment after 18 and 24 hours. Arcs link highly significant interactions (q-value $< 5 \times 10^{-5}$) and arc widths are proportional to the normalised number of reads recorded for the interaction (Methods **??**).

Table 1: 4C sequencing library statistics. 4C experiments are summarised as total number of reads in each experiment and the percentage of those reads labelled "duplicates". Note in 4C these duplicates are not artifactual and instead result from large numbers of contacts nearby to the viewpoint.

	TSA-		TSA+	
	MlucI	HaeIII	18h	24h
Reads (Million)	10	10.4	42.2	24.2
Duplicated (%)	62.8	74.0	71.2	84.4

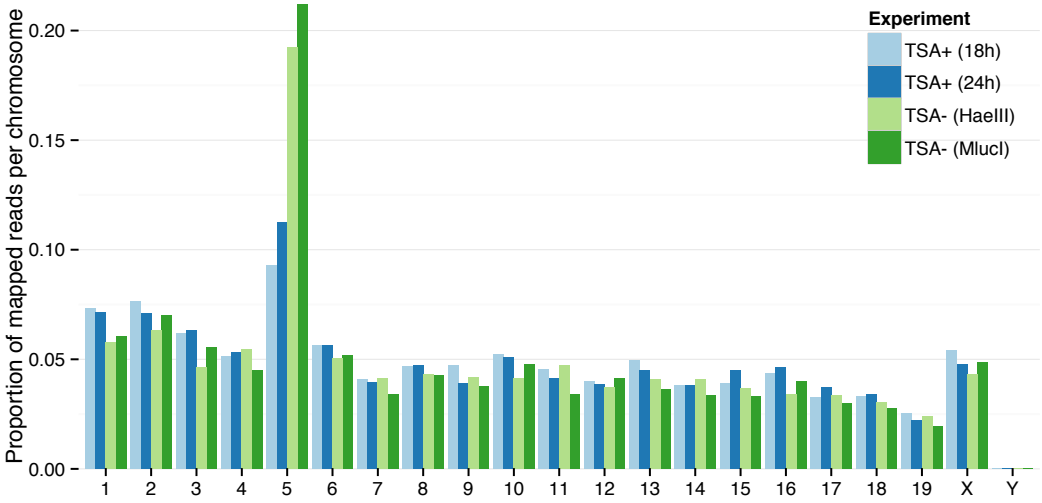

Figure 4: The bait chromosome is enriched for 4C sequencing reads. Chromosome 5 is visibly enriched for 4C reads as it contains the ZRS bait region (or viewpoint). Untreated control samples (*TSA-*) were assayed with two different secondary restriction enzymes (4-cutters HaeIII and MlucI).

to 84.4%. This suggests that while the assay does appear to be working, there may be extraneous noise and non-bait interactions in the sequencing library.

We found the proportion of reads mapped to the fair region chromosome, chromosome 5 in this case, fell below the prescribed level of 50%. Looking at three experiments (two from the untreated control), we find instead that between approximately 10–20% of all reads mapped to the bait chromosome (Fig. 4). While this is still a clear enrichment over non-bait chromosomes, it suggests the assay results suffer from either increased *trans*-contact noise or decreased *cis*-contact enrichment around the bait region.

Lower than expected levels of both sequence duplication and bait chromosome enrichment suggest loss of signal around the bait region itself. This is the area where we'd expect both very high levels of duplication (identical restriction fragment pairings between nearby genomic locations) and a majority of all sequencing reads, driving the overall chromosome enrichment. The precise reason for the discrepancy is unclear but suggests the results may have a lower signal-to-noise ratio than has previously been achievable in 4C experiments.[8] Potentially the signal-to-noise ratio could be

improved by utilising a double cross-linking procedure such as that used in Lin *et al.*[13]

1.3 3D MODELLING OF CHROMATIN FIBRE

Chromosome conformation capture allow investigation of genome organisation, but such data are commonly analysed using one or two-dimensional representations. A growing set of algorithms looks instead to rebuild the three-dimensional polymer trajectory of a chromatin fibre, using Hi-C or 5C data as input (e.g. 14–21). Intuitively, in each method the interaction frequency between two regions is idealised as inversely proportional to their physical distance (where possible and according to various other constraints). Where these methods differ is in their approaches to solving this optimisation problem. We chose the `AutoChrom3D` method[19] for use in this work (described in Section 1.3.1) as the algorithm can accept 5C input and model polymers at high resolution of 8 kb or greater.

1.3.1 `AutoChrom3D` method

The procedure implemented in `AutoChrom3D` can be summarised as:[19]

1. The chromatin fibre is represented as beads-on-a-string, with $N_{beads} = \lceil \frac{L}{R} \rceil$ (where L is the length of the region and R the resolution)
2. A local compaction parameter is calculated using a sliding window of each 50 adjacent beads (intra-window contacts are averaged and compared to those over the whole region under study)
3. Interaction frequency between beads of a given genomic distance is modelled as a Poisson-distributed random variable and noisy or unstable contacts, considered in the context of neighbouring beads, are filtered
4. This filtered set of interaction frequencies are then normalised using the previously-calculated compaction parameter to give an $N_{beads} \times N_{beads}$ matrix of interaction strengths
5. Interaction strength is converted to spatial distance through two linear transformations based on experimental observations of nuclear occupancy and regional flexibility[22]
6. Cartesian co-ordinates are then calculated via non-linear constrained optimisation of pairwise spatial distances using LINGO[23]

Table 2: Measurement distances between ZRS and SHH in each inferred 3D structure. Distances are given in arbitrary units. *Shh* spans two beads of the polymer model, hence two distances are calculated in each case (d_1, d_2). RMSD is the minimised root mean squared deviation between the two structures and is given as a relative unitless quantity. The radius of gyration (gyradius) is also shown.

		Distance		RMSD	**Gyradius** (μm)	
		TSA-	TSA+		TSA-	TSA+
88fp	d_1	5.4	5.1	1.701	0.244	0.244
	d_2	4.1	3.9			
MD	d_1	6.2	3.3	2.377	0.217	0.205
	d_2	4.8	2.0			

1.3.2 Modelling the *Shh* region with 5C

5C data was generated by our collaborators over the same *Shh*–ZRS region as was assayed with 4C (Fig. 1; Section 1.2) with the aim of developing a multi-point perspective on local chromatin conformation beyond that available from 4C data.

We used this 5C experimental data in combination with a particular three-dimensional inference program (`AutoChrom3D`[19]) in an attempt to compare polymer trajectories in TSA treated and untreated 88fp mouse cells, a similar and complimentary cell line to that used in earlier 4C experiments (14fp). As a control, 5C was also performed on mandibular (MD) cells, with and without TSA treatment, which do not express *Shh*. Prior to structural modelling, the `my5C` program was used to generate normalised 5C interaction frequencies.[24]

We find that TSA treatment of 88fp cells does appear to slightly reduce the distance between *Shh* and ZRS in inferred 3D structures (Fig. 5), however this difference is overshadowed—to our surprise—by that observed in the non-expressing MD cell line. This latter mandibular cell line undergoes a large structural transition which brings together the *Shh* gene and the ZRS enhancer. Measurements between these elements for each structure are shown in Table 2.

We also report a greater overall structural shift following TSA treatment in the MD cell line, with an RMSD between the two structures of 2.377 arbitrary units, relative to 1.701 between TSA+ and TSA- 88fp cells. The radius of gyration, unchanged in 88fp, is also decreased in the MD cell line following TSA treatment, indicating the region becomes more compact following TSA treatment (Table 2).

1.3.3 Repeat simulations

We have shown what appears to be a structural shift in the *Shh*-ZRS locus per 3D modelling predictions (Section 1.3.2). It is of interest to assess the stability and reproducibility of these results through repeat simulations of the polymer trajectory. At this point it is unclear whether the *Shh*-ZRS bound state represents a firm consensus

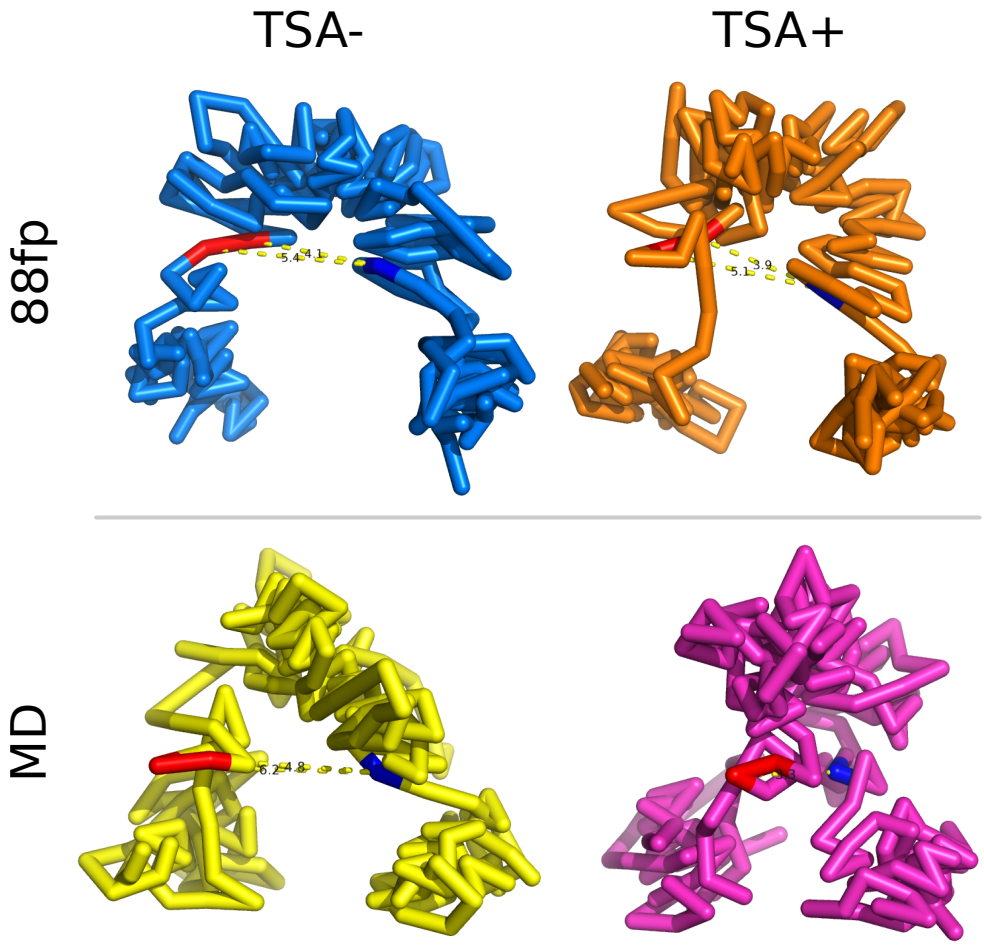

Figure 5: Inferred polymer trajectories of the ZRS–SHH region following TSA treatment in two cell lines. 3D structures are shown for 5C experiments assaying the region around *Shh* (*red*) and ZRS (*blue*) in an *Shh*-expressing limb bud cell line (88fp) and a non-expressing mandibular cell line (MD). Labelled measurements are given in Table 2. Structures were predicted by `AutoChrom3D`[19] using 210 × 8 kb beads per polymer.

over the cell population, or an alternative structure with similar optimisation energy to that of the more non-bound state.

We re-ran simulations of the 3D chromatin fibre in the *Shh*-ZRS region a total of five times (Fig. 6). In every case, the algorithm generates the known *Shh*-ZRS TAD as a compacted domain bookended by the two loci under study. This sanity check ensures the results are broadly compatible with our *a prior* expectation of the region's structure given the 2D heatmap representation of 5C data (Fig. 1).

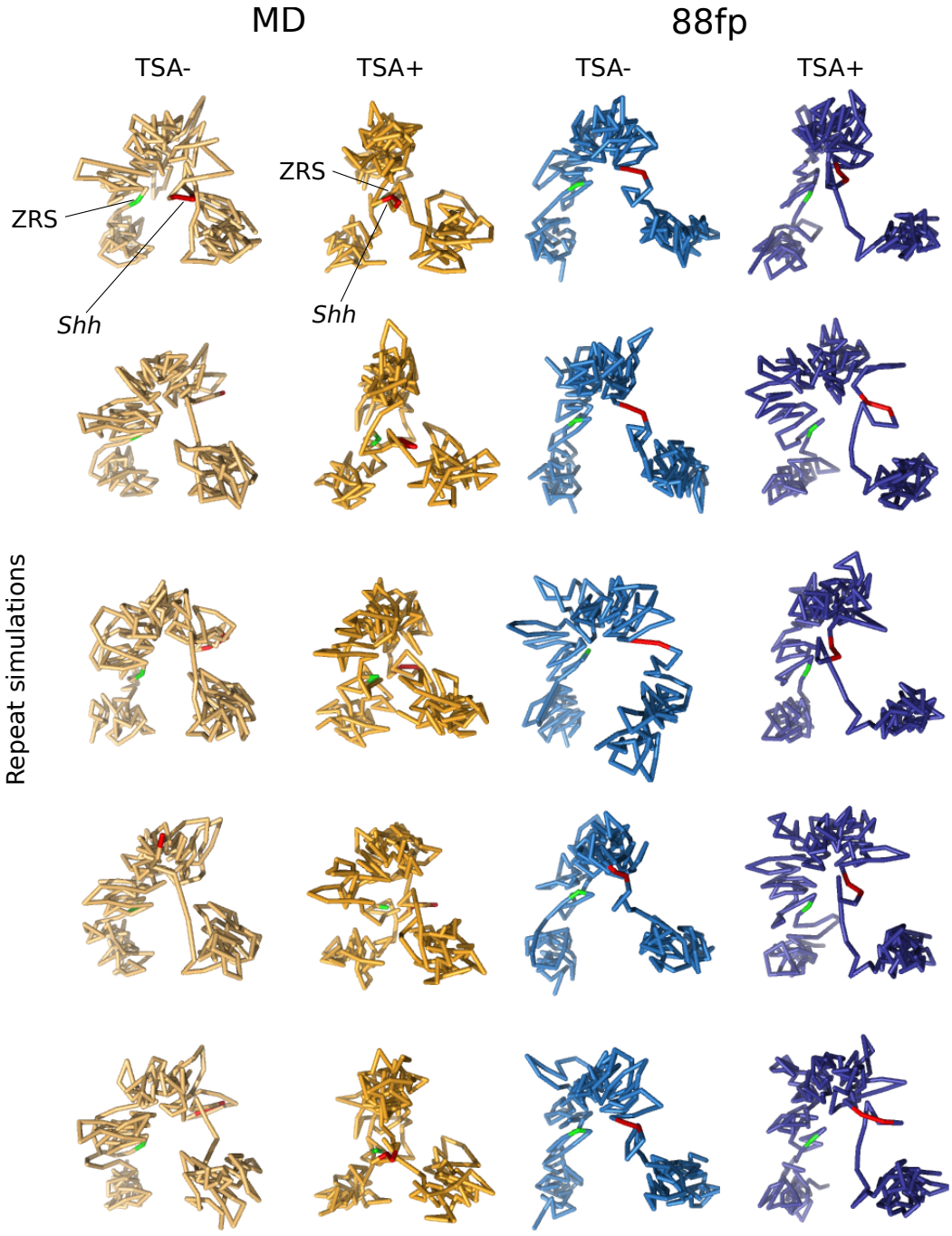

Figure 6: Repeat simulations of 3D polymer trajectories in the *Shh*–ZRS region. 3D structures are shown for 5C experiments assaying the region around *Shh* (*red*) and ZRS (*blue*) in an *Shh*-expressing limb bud cell line (88fp) and a non-expressing mandibular cell line (MD). Structures were aligned as whole molecules with the uppermost replicate in each column.

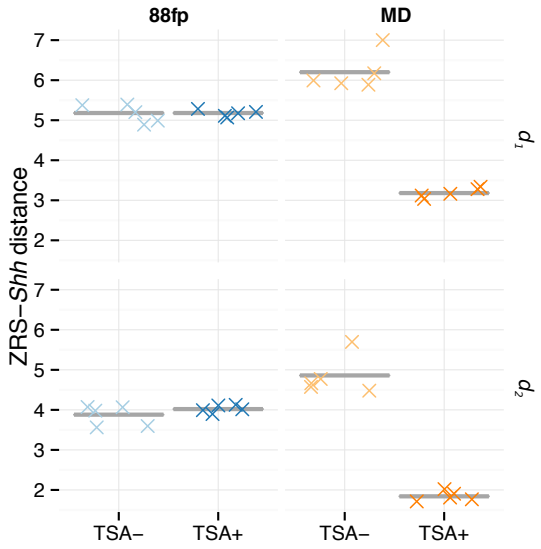

Figure 7: *Shh*–ZRS distance measurements from repeated 3D polymer simulations. Measurements were taken from 5 replicate 3D simulations (shown in Fig. 6). Distances are given in arbitrary units. *Shh* spans two beads of the polymer model, hence two distances are calculated in each case (d_1, d_2).

REFERENCES

[1] Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**(7398): 376–80.

[2] Anderson E, Devenney PS, Hill RE, Lettice La (2014) Mapping the Shh long-range regulatory domain. *Development (Cambridge, England)*, (September): 1–10.

[3] Anderson E, Peluso S, Lettice La, Hill RE (2012) Human limb abnormalities caused by disruption of hedgehog signaling. *Trends in Genetics*, **28**(8): 364–373.

[4] Hill RE, Lettice La, B PTRS (2013) Alterations to the remote control of Shh gene expression cause congenital abnormalities Alterations to the remote control of Shh gene expression cause congenital abnormalities Author for correspondence :. (May).

[5] Laurell T, Vandermeer JE, Wenger AM, Grigelioniene G, Nordenskjöld A, Arner M, Ekblom AG, Bejerano G, Ahituv N, Nordgren A (2012) A novel 13 base pair insertion in the sonic hedgehog ZRS limb enhancer (ZRS/LMBR1) causes preaxial polydactyly with triphalangeal thumb. *Human Mutation*, **33**(7): 1063–1066.

[6] Lettice La, Hill AE, Devenney PS, Hill RE (2008) Point mutations in a distant sonic hedgehog cis-regulator generate a variable regulatory output responsible for preaxial polydactyly. *Human Molecular Genetics*, **17**(7): 978–985.

[7] Bouwman BA, de Laat W (2015) Getting the genome in shape: the formation of loops, domains and compartments. *Genome Biology*, **16**(1): 154.

[8] Stadhouders R, Kolovos P, Brouwer R, Zuin J, van den Heuvel A, Kockx C, Palstra RJ, Wendt KS, Grosveld F, *et al.* (2013) Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions. *Nature protocols*, **8**(3): 509–24.

[9] Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**(1): 10–12.

[10] Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature methods*, **9**(4): 357–9.

[11] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**(16): 2078–2079.

[12] Andrews S (2015) FastQC (v0.10.1): A quality control tool for high throughput sequence data.

[13] Lin YC, Benner C, Mansson R, Heinz S, Miyazaki K, Miyazaki M, Chandra V, Bossen C, Glass CK, Murre C (2012) Global changes in the nuclear positioning of genes and intra- and interdomain genomic interactions that orchestrate B cell fate. *Nature immunology*, **13**(12): 1196–204.

[14] Baù D, Sanyal A, Lajoie BR, Capriotti E, Byron M, Lawrence JB, Dekker J, Marti-Renom Ma (2011) The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. *Nature structural & molecular biology*, **18**(1): 107–14.

[15] Hu M, Deng K, Qin Z, Dixon J, Selvaraj S, Fang J, Ren B, Liu JS (2013) Bayesian inference of spatial organizations of chromosomes. *PLoS computational biology*, **9**(1): e1002893.

[16] Varoquaux N, Ay F, Noble WS, Vert Jp (2014) A statistical approach for inferring the 3D structure of the genome. *Bioinformatics (Oxford, England)*, **30**(12): i26–i33.

[17] Lesne A, Riposo J, Roger P, Cournac A, Mozziconacci J (2014) 3D genome reconstruction from chromosomal contacts. *Nature methods*, (september).

[18] Trieu T, Cheng J (2014) Large-scale reconstruction of 3D structures of human chromosomes from chromosomal contact data. *Nucleic Acids Research*, **42**(7): 1–11.

[19] Peng C, Fu LY, Dong PF, Deng ZL, Li JX, Wang XT, Zhang HY (2013) The sequencing bias relaxed characteristics of Hi-C derived data and implications for chromatin 3D modeling. *Nucleic Acids Research*, **41**(19).

[20] Ay F, Bunnik EM, Varoquaux N, Bol SM, Prudhomme J, Vert JP, Noble WS, Le Roch KG (2014) Three-dimensional modeling of the P. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. *Genome research*.

[21] Caudai C, Salerno E, Zoppè M, Tonazzini A (2015) Inferring 3D chromatin structure using a multiscale approach based on quaternions. *BMC Bioinformatics*, **16**(1): 234.

[22] Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L (2012) Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature biotechnology*, **30**(1): 90–8.

[23] LINDO (2015) LINGO (v15.0): Optimization modeling software for linear, nonlinear, and integer programming.

[24] Lajoie BR, van Berkum NL, Sanyal A, Dekker J (2009) My5C: web tools for chromosome conformation capture studies. *Nature methods*, **6**(10): 690–691.