Forum

Beyond the Linear Genome: Paired-End Sequencing as a Biophysical Tool

Viviana I. Risca[1] and William J. Greenleaf[1,2,*]

Paired-end sequencing has enabled a variety of new methods for high-throughput interrogation of both genome structure and chromatin architecture. Here, we discuss how the paired-end paradigm can be used to interpret sequencing data as biophysical measurements of *in vivo* chromatin structure that report on single molecules in single cells.

Paired-End Sequencing

Pairwise correlations are perhaps the simplest and yet most powerful measurements in biology. Correlation measurements are fundamentally enabled by pairwise measurements of variables that share a fixed characteristic. Pairs of measurements can link biological elements in space (e.g., the relaxation of two proximal atoms exhibiting spin coupling), time (e.g., correlations of neuronal firing), or function (e.g., two mutations capable of compensating for each other), allowing reconstruction of biological components from macromolecular structures to complex biological signaling networks. Pairwise measurements become still more informative when deployed in high throughput to comprehensively map interactions in a biological system. Here, we explore the ways in which pairwise measurements can be made using DNA sequencing-based assays, which effectively report single-molecule information, allowing multiplex biophysical measurements in living cells.

Modern DNA sequencing technology routinely produces hundreds of millions of short reads spanning tens to hundreds of base pairs for only a few thousand dollars. From its earliest and simplest application to reading out genome sequences, DNA sequencing has evolved, through the generation of diverse assays that use short DNA fragments as a readout, into a powerful tool for cell biology and nucleic acid biophysics, enabling assays of protein–DNA interactions [1], RNA expression and splicing [1–3], and ribosome–RNA interactions [4]. These methods are complementary to established lower-throughput assays such as live-cell and immunofluorescence microscopy, which, although lower in throughput, can often access temporal dynamics that sequencing cannot and can validate observations from sequencing experiments. Although a thorough review of the many applications of sequencing to biophysical measurements is beyond the scope of this forum article, we will focus on methods that allow for correlated measurements using paired-end sequencing, a modality that is particularly promising for maximizing biophysical and cell biological insight.

In most applications, paired-end sequencing is carried out by performing two (or more) sequential rounds of sequencing-by-synthesis on each library molecule (Figure 1A), and these separate reads are identified as linked in subsequent analysis. For libraries in which the insert size of genomic DNA exceeds the length of each read (which is often true for short-read platforms), reading both ends of each insert allows mapping of the fragment onto a reference genome and determination of the insert length. If reads align discordantly (i.e., if fragment lengths exceed the known size range of the library, or if orientations are inconsistent), this information can be used to infer structural variation of the sequenced genome. Alternately, single-read methods can be applied to libraries that are circularized, creating 'mate pairs' representing inserts of several kilobases [5] (Figure 1B). Members of a library can also be cleaved and ligated consecutively as paired-end tags (PETs) [6], then read out on a single-end platform (Figure 1C).

Measurements of Molecular Contiguity

Paired-end sequencing has been extensively applied to measure the contiguity of single DNA molecules, a crucial step in *de novo* genome sequence assembly, haplotype phasing, and the detection of structural variation (Figure 2A) [5,7]. Recent extensions of these ideas have combined whole-genome amplification with paired-end sequencing to detect the emergence of chromosome rearrangements in single cells of a human embryo over a single cell cycle, enabling observation of pairs of daughter cells with reciprocal rearrangements [8]. Much longer contiguity measurements can be made with CPT-seq, a technique that combines transposase-based linking of molecules, multiple rounds of barcoding, and paired-end barcode reads identifying linked molecules [5]. Measurements of molecular contiguity have also been applied to RNA, using PETs or mate pairs made via circularization to simultaneously sequence the 5′ and 3′ ends of single transcripts, and paired-end data is used by many algorithms to improve the detection of RNA splicing variants [2] (Figure 2B). All of these methods constitute direct measurements of individual DNA or RNA molecules, making it possible, especially in single-cell assays, to quantitatively study chromosome recombination, the mechanisms that preserve the integrity of the genome, and RNA splicing with ultimate sensitivity.

Mapping Genome Architecture by Proximity Ligation and Paired-End Sequencing

Proximity ligation-based measurements of 3D chromosome conformation, including chromosome conformation capture (3C), its high-throughput variant, Hi-C [9], and chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) [6] also critically rely on paired-end sequencing (Figure 2C). The frequency of observation

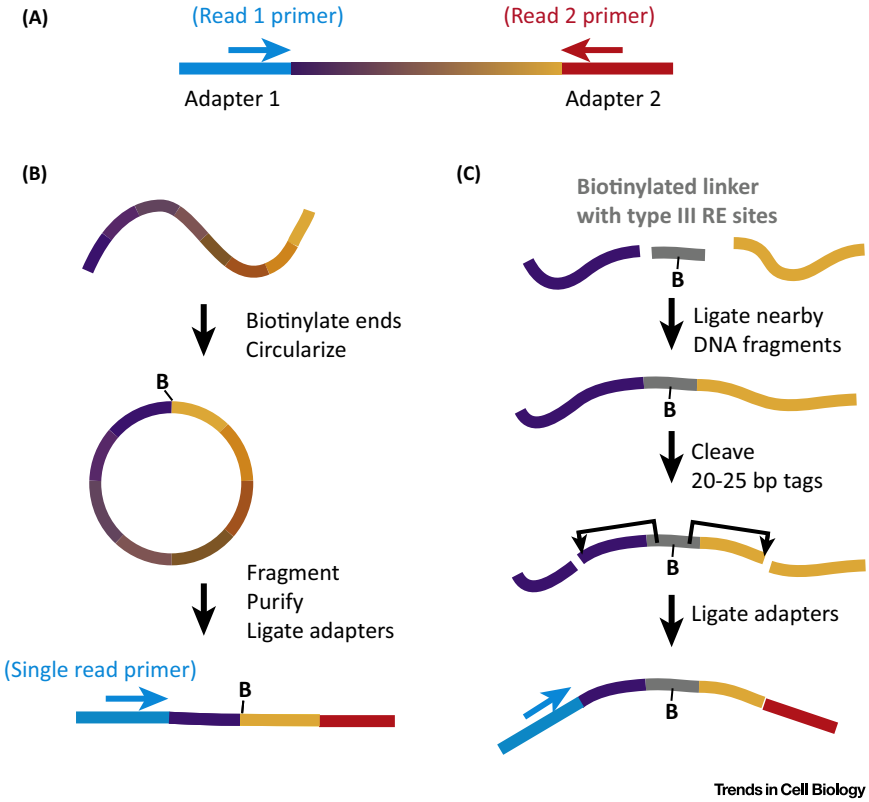

Figure 1. Types of Paired-End Libraries. (A) Simple library molecule with two adapter sequences complementary to two read primer sequences for simple paired-end sequencing. (B) Example of mate-pair library production from long (several kilobase) fragments of genomic DNA. (C) Example of paired-end tag library generation, in which a Type III restriction endonuclease (RE) with recognition sites encoded in the common linker is used to cleave 20–25 bp sequence tags from genomic DNA flanking the linker [6].

of chimeric junctions produced by proximity ligation provides a measurement of the 3D proximity of the two loci (after accounting for biases attributable to factors such as DNA fragment length) [9]. Hi-C has revealed that the genome is organized into megabase-scale topologically associating domains (TADs) within which 3D associations occur with high frequency, as compared with inter-TAD associations (Figure 2C) [9]. ChIA-PET (Figure 2C) has allowed finer-grained, cell type-specific, and stimulus-specific identification of 3D spatial correlations between loci, such as enhancers, and coregulated gene promoters, which appear to also temporally correlate with the induction of gene expression [6].

Although much proximity ligation work has been descriptive, ligation frequencies are biophysical measurements that map intermolecular distances and promise to become an increasingly useful and quantitative tool for both geneticists and cell biologists interested in chromosome structure and dynamics. Before proximity ligation can be calibrated precisely to spatial distance, several caveats remain to be considered and further studied. First, because cells are crosslinked before DNA fragmentation and proximity ligation, nuclear structure may be distorted at the nanometer scale, leading to apparent contacts between molecules that are in fact hundreds of nanometers apart in the native cell, but may be associated with the same intracellular structure [10]. Second, it is not yet clear to what extent proximity ligation events are capturing rare fluctuations rather than reporting on a stable conformation, although fluorescence *in situ* hybridization (FISH) data do, in most (but not all) cases, corroborate Hi-C data [9]. Early hints from high-resolution FISH combined with modeling suggest that proximity ligation maps are an average of highly dynamic, fluctuating chromosome conformations [11]. Individual chimeric reads may therefore be interpreted not just as measurements of proximity between two loci on a single DNA molecule but also single-cell and single-time slice measurements from a fluctuating ensemble of conformations.

Paired-End Sequencing for Fine-Scale Chromatin Structure Assays

At much higher spatial resolution, paired-end sequencing has proven useful in high-resolution mapping of the fine-grained structure of chromatin that consists of nucleosome positions and DNA-bound transcription factors. DNA-bound proteins can be mapped with 'footprinting', in which a nuclease is used to digest free DNA, while leaving intact any fragments that are protected by a bound transcription factor or nucleosome (Figure 2D). Fragments generated by footprinting are the result of two cleavage events that must have occurred in the same cell and on the same DNA molecule, in the same open chromatin region or flanking the same DNA-bound protein. Paired-end footprinting therefore encodes correlations in the local chromatin state between the two ends of any DNA fragment.

Although footprint data can and have been analyzed using single-end sequencing, the correlation information encoded in paired-end reads provides several advantages: (i) read pairs can be filtered for the exact size of the complex of interest, excluding background reads, and increasing the resolution of inferences [1,12]; (ii) fragment size can be used to differentiate footprints arising from different classes of DNA-bound elements, that is, transcription factors versus nucleosomes, and analyze them separately [1]; (iii) a complete picture of chromatin

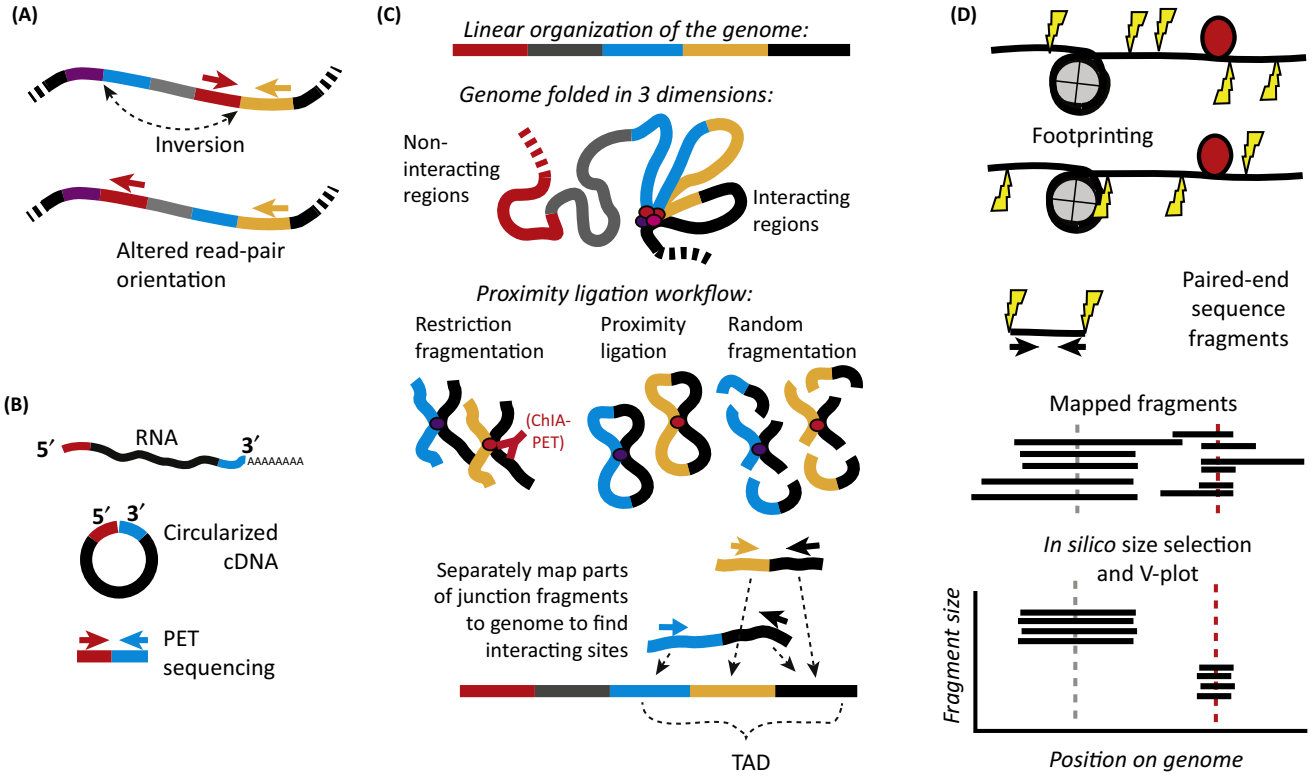

Figure 2. Extracting Biophysical and Structural Information from Mapped Paired-End Reads. (A) Paired-end or mate-paired read orientation can be used to detect structural variation in genomes, such as the inversion shown here. (B) RNA 5′ and 3′ ends can be simultaneously mapped with RNA-PET, an example of use of paired-end sequencing in transcriptomics. (C) Hi-C and ChIA-PET (chromatin interaction analysis by paired-end tag sequencing) report on 3D contacts of loci along the linear genome (shown as colored blocks). In Hi-C [9], crosslinked and lightly permeabilized nuclei are treated with restriction endonucleases to cut the genome into short pieces spanning several hundred base pairs. The ends of spatially proximal DNA fragments held together by crosslinked proteins (red or purple circles) are then ligated to generate chimeric molecules (blue–black and yellow–black). Paired-end sequencing of these chimeric molecules yields pairs of reads that can be separately mapped to the genome, with each read pair reporting contact between a pair of loci in a single cell (e.g., blue–black, yellow–black). A region with a high density of self-contacts (blue/yellow/black) is designated a topologically associating domain (TAD). ChIA-PET [6] operates similarly, but includes immunoprecipitation (red Y) to isolate interactions mediated by a particular protein of interest (red oval). (D) Footprinting assays use cleavage of unprotected DNA by an enzyme (lightning bolt) to generate short DNA fragments. Paired-end sequencing of these fragments provides precise fragment length information that can improve signal-to-noise and permit the footprinting of multiple classes of particles with V-plots of fragment length versus genomic coordinates [1].

structure showing the locations of nucleosomes and transcription factors at any locus of interest can be analyzed using a V-plot, which plots fragment lengths versus fragment center distances from the locus (Figure 2D) [1]; and (iv) paired-read alignment makes it possible to distinguish reads mapping to identical coordinates based on the different coordinates of their paired reads, allowing greater dynamic range after the removal of amplification-generated duplicates. Although the detailed characteristics of data produced by footprinting methods vary based on the enzyme used, they can all, in principle, be used to determine both the position and occupancy of DNA-bound nucleosomes and transcription factors, producing an *in vivo* measurement of binding affinity as well as kinetics with sufficient time resolution to distinguish populations of molecules. Caveats to interpretation of occupancy measures arise from such factors as fragment size biases in both PCR and sequencing as well as in the substantial sequence bias of the enzyme used for footprinting, which may give rise to artefactual apparent footprints [13]. In paired-end analysis, this bias must be accounted for jointly, because the frequency that a particular cleavage event is observed depends on the resulting fragment being observed, which in turn depends on a second cleavage occurring within a given distance. However, with accurate background models of sequence and length bias, the full potential of paired-end chromatin footprinting analysis for genome-wide quantitative understanding of chromatin architecture may be realized.

Concluding Remarks

As sequencing finds more applications in cell biology, paired-end approaches provide a powerful paradigm for thinking in

terms of single DNA fragments with ends that report on correlated DNA accessibility, spatial proximity between genomically distant loci, or splicing and rearrangement events. Calibration is a major challenge in integrating sequencing-based approaches with existing cell biological approaches, such as FISH or biochemical approaches, for example, *in vitro* measurements of transcription factor affinity. However, opportunities exist to map the correspondence between sequencing-based and classical measurements of distance or affinity in a cell, bringing these high-throughput methods into the quantitative realm. Together, such complementary approaches hold great promise for a quantitative understanding of genome and chromosome biology.

Acknowledgments

W.J.G. acknowledges support from National Institutes of Health (NIH) grants R21HG007726, U01HG007919, U1AI057266, and P50HG007735, the Rita Allen Foundation and the Human Frontier Science Program. V.I.R. acknowledges support from the Walter V. and Idun Berry Postdoctoral Fellowship and the Stanford Center for Systems Biology Seed Grant. We apologize to our colleagues whose work could not be cited owing to length constraints.

Disclaimer Statement

W.J.G. is named as an inventor on a patent application filed by Stanford University regarding ATAC-seq technique, and is a scientific co-founder of Epinomics.

[1]Department of Genetics, Stanford University School of Medicine, Stanford, CA, 94305, USA
[2]Department of Applied Physics, Stanford University, Stanford, CA, 94305, USA

*Correspondence: wjg@stanford.edu (W.J. Greenleaf).

http://dx.doi.org/10.1016/j.tcb.2015.08.004

References

1. Zentner, G.E. and Henikoff, S. (2014) High-resolution digital profiling of the epigenome. *Nat. Rev. Genet.* 15, 814–827
2. de Klerk, E. *et al.* (2014) RNA sequencing: from tag-based profiling to resolving complete transcript structure. *Cell. Mol. Life Sci.* 71, 3537–3551
3. Mayer, A. *et al.* (2015) Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. *Cell* 161, 541–554
4. Ingolia, N.T. (2014) Ribosome profiling: new views of translation, from single codons to genome scale. *Nat. Rev. Genet.* 15, 205–213
5. Adey, A. *et al.* (2014) In vitro, long-range sequence information for de novo genome assembly via transposase contiguity. *Genome Res.* 24, 2041–2049
6. Fullwood, M.J. *et al.* (2009) An oestrogen-receptor-α-bound human chromatin interactome. *Nature* 462, 58–64
7. Amini, S. *et al.* (2014) Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat. Genet.* 46, 1343–1349
8. Voet, T. *et al.* (2013) Single-cell paired-end genome sequencing reveals structural variation per cell cycle. *Nucleic Acids Res.* 41, 6119–6138
9. Rao, S.S.P. *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159, 1665–1680
10. Belmont, A.S. (2014) Large-scale chromatin organization: the good, the surprising, and the still perplexing. *Curr. Opin. Cell Biol.* 26, 69–78
11. Giorgetti, L. *et al.* (2014) Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell* 157, 950–963
12. Gaffney, D.J. *et al.* (2012) Controls of nucleosome positioning in the human genome. *PLoS Genet.* 8, e1003036
13. Meyer, C.A. and Liu, X.S. (2014) Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat. Rev. Genet.* 15, 709–721