Core promoters across the genome

Promoter strength is measured genome-wide with two high-throughput reporter assays.

Nevena Cvetesic and Boris Lenhard

Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, London, UK; MRC London Institute of Medical Sciences, London, UK

Promoters, the sites that define where transcription starts, are fundamental to genome function. Although methods have been developed to map promoter activity genome-wide, from mature RNA (e.g. CAGE4 and CapSeq[5]) and recently from nascent transcripts (e.g. GRO-cap[6]), we still lack a systematic understanding of the sequence requirements for promoter activity because existing approaches measure the integrated effect of everything that influences promoter activity, including proximal and distal enhancers, other sequence features, and chromatin environment and spatial organization. Thus, the specific contribution of promoter sequences themselves is unknown. Two reports in this issue, by Arnold *et al.*[7] and van Arensbergen *et al.*[8], now present methods for measuring the autonomous promoter activity of random sequences genome-wide. These methods also identify "cryptic" sequences that can activate transcription in the reporter setting but are prevented from doing so in their native genome context. In addition, the system of Arnold *et al.*[7] detects promoter sequences that respond (non-autonomously) to an enhancer.

Promoters are recognized by proteins in the transcription initiation complex, and in many cases are regulated by context-specific regulatory elements, both proximal and distal. Core promoters encompass roughly 40 bp upstream and downstream of the transcription

starts sites and represents a physical place where transcription initiation machinery (general transcription factors, GTFs) binds to direct initiation by RNA polymerase II. They have multiple, functionally distinct[1] and sometimes physically overlapping[2] architectures, which determine their responsiveness to long-range regulation and to developmental and differentiation regulatory programs (reviewed in ref. 3).

Methods to study promoter activity genome-wide include CAGE[4] and CapSeq[5], which measure mature accumulated RNA. Both methods select only intact/capped mRNA either by biotinylation of the cap and affinity purification using streptavidin resin (CAGE), or by enzymatic treatment with CIP and TAP to leave a ligatable 5'end only on capped mRNA. A more recent method called GRO-cap[6] allows detection of transcription start sites from nascent transcripts. First, a nuclear run-on assay is performed to select nascent RNAs, followed by removal of the uncapped RNA through enzymatic treatment. However, none of these approaches can deconvolute the effects of promoter sequences from those of other genomic features. This is now possible using either of the new methods---self-transcribing active core promoter-sequencing (STAP-seq) by Arnold *et al.*[7] (**Fig. 1a**) and survey of regulatory elements (SuRE) by van Arensbergen *et al.*[8] (**Fig. 1b**).

STAP-seq is a clever inverse of the successful STARR-seq method[9] for probing enhancer activity developed by the same group. The authors construct a candidate library - and use it to assess sequence-intrinsic core promoter strength and promoter potential of candidate regions that span the entire *Drosophila* genome (Figure 1a). The libraries are constructed by shearing the genome and cloning selected ~200-bp-long fragments into reporter constructs at a position appropriate for a core promoter---upstream of a protein-coding open reading frame and downstream of a defined enhancer. The reporter libraries are then transfected into target cell lines, and gene expression is measured (see below). In parallel, an enhancer-less reporter construct is used to assess basal core promoter activity.

Thus, the experiments reveal genomic sequences with both enhancer-driven and enhancer-independent promoter activity. Transcriptional output is measured with a modified 5'-tag sequencing method (CapSeq[5]) for transcription start site mapping, allowing single-nucleotide-resolution mapping of transcription initiation sites. The authors demonstrate STAP-seq in a study of sequences spanning the entire *Drosophila* genome.

In the second paper, van Arensbergen *et al.*[8] introduce SuRE, which similarly queries the autonomous promoter activity of random genomic fragments outside of their genomic context using a defined reporter assay. One important difference relative to STAP-seq is the size of the tested fragments: to allow the entire human genome to be probed at high coverage, the authors use a wide fragment size range of 0.2-2 kbp. The fragments are cloned into a promoter- and enhancer-less reporter upstream of a barcode sequence, an open reading frame and a polyadenylation signal. The library is then transfected into a test cell line, in this case **K562 cells**, and transcriptional output is measured by RNA-seq. An intermediate experimental step is necessary to link the tested genomic fragments with the transcriptional output through association of barcodes and fragments in the reporter library. The output signal is normalized by the frequency of barcodes detected in the plasmid library. While SuRE signal measures autonomous transcriptional initiation of genomic sequences, GRO-cap measures their endogenous initiation in the genome. The authors use the ratio of the two as a measure of promoter autonomy.

A limitation of SuRE compared to STAP-seq is lower resolution: the peaks obtained correspond in size to the tested fragments, whereas STAP-seq maps initiation start sites at single-nucleotide resolution. However, because the tested fragments overlap, van

Arensbergen *et al.*[8] are able to computationally enhance the resolution of SuRE signals by determining the minimal regions essential for reporter expression.

Although both STAP-seq and SuRE succeed in interrogating the promoter potential of random genomic fragments, the methods have a few limitations. First, transcriptional output measurements are standardized by using reporter constructs, which diverge from the endogenous genomic context. Second, in both approaches, enhancers may be situated very close to core promoters, raising the question of how to define autonomous, basal promoter activity, especially when large regions are tested as in SuRE. Recent work in the field demonstrates that a promoter's transcriptional output depends on the position of the enhancer.

Nonetheless, the two studies yield some very interesting results. van Arensbergen *et al.*[8] focus on exploring promoter autonomy, which they define as the ratio of autonomous to endogenous transcription initiation (*i.e.,* the ratio of SuRE to GRO-cap[5] signals). Although the trends are noisy, promoter autonomy is overall greater in ubiquitously expressed genes that have a lower dependence on the number of endogenous enhancers attributed to them. The same sequence subregions (parts of tested fragments determined by generalized linear modeling to contribute most to the autonomous transcription) govern both sense and antisense autonomous transcription. This implies that sense core promoter can also be responsible for antisense transcription, i.e. that a separate antisense core promoter is not necessary for antisense initiation.

Arnold *et al.*[7] provide the first comprehensive assay of promoter responsiveness to enhancer regulation. They call this responsiveness promoter strength, and define it as the normalized ratio of transcriptional output in the presence and absence of an enhancer. Their analysis of the data identifies the strength and relative position of core promoter motifs (short sequences such as TATA box or Initiator (Inr), which are recognized by DNA-binding proteins involved in initiating transcription) as the major determinants of overall promoter strength, which explains the wide dynamic range of achievable promoter strength through combinatorial changes. Housekeeping promoters show a much reduced dynamic range in their strength compared to core promoters of developmentally regulated and tissue-specific genes, with the latter showing a staggering 1000-fold dynamic range in responsiveness to enhancers. Moreover, the strength of a core promoter correlates with the number of endogenous enhancers around it, a result that complements the findings from SuRE establishing that autonomy anti-correlates with the number of the annotated endogenous enhancers.

These first results generated by STAP-seq and SuRE strengthen the case for the core promoter as a genomic element at the center of transcriptional regulation. The two papers only scratch the surface of what can be investigated by analyzing the data they generated. For example, the SuRE study not only confirms that many known enhancers and some repetitive elements possess autonomous promoter activity but also reveals that many sequences with no promoter activity in their native genomic context show promoter activity reproducibly in the SuRE assay. Such 'cryptic' promoters raise questions about how they are kept silent *in vivo* and whether they might be active under certain conditions. The STAP-seq study does not address the phenomenon of cryptic promoters, but it shows that different (developmental and housekeeping) enhancers can be used to detect different classes of promoters genome-wide, a

potentially fruitful avenue for future. More generally, the data produced in the two reports could be mined to investigate promoter properties that were not examined, such as the precision of endogenous *vs* autonomous transcription initiation, or differences in motif content between promoters that exhibit greater autonomy or enhancer responsiveness/promoter strength. Integrated analysis of these and other datasets will surely enable further rapid progress in characterizing different promoter architectures and their sequence determinants.

1. Zabidi, M. A. *et al.* Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518,** 556–559 (2015).
2. Haberle, V. *et al.* Two independent transcription initiation codes overlap on vertebrate core promoters. *Nature* **507,** 381–385 (2014).
3. Haberle, V. & Lenhard, B. Promoter architectures and developmental gene regulation. *Seminars in Cell & Developmental Biology* **57,** 11–23 (2016).
4. Kodzius, R. *et al.* CAGE: cap analysis of gene expression. *Nat Meth* **3,** 211–222 (2006).
5. Gu, W. *et al.* CapSeq and CIP-TAP identify Pol II start sites and reveal capped small RNAs as C. elegans piRNA precursors. *Cell* **151,** 1488–1500 (2012).
6. Core, L. J. *et al.* Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* **46,** 1311–1320 (2014).
7. Arnold, C. D. *et al.* The genome-wide core promoter activity assay STAP-seq reveals a wide range of sequence-intrinsic core promoter strengths. *Nat Biotechnol* (2017).
8. van Arensbergen, J. *et al.* Genome-wide mapping of autonomous promoter activity in human cells. *Nat Biotechnol* (2017).
9. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339,** 1074–1077 (2013).

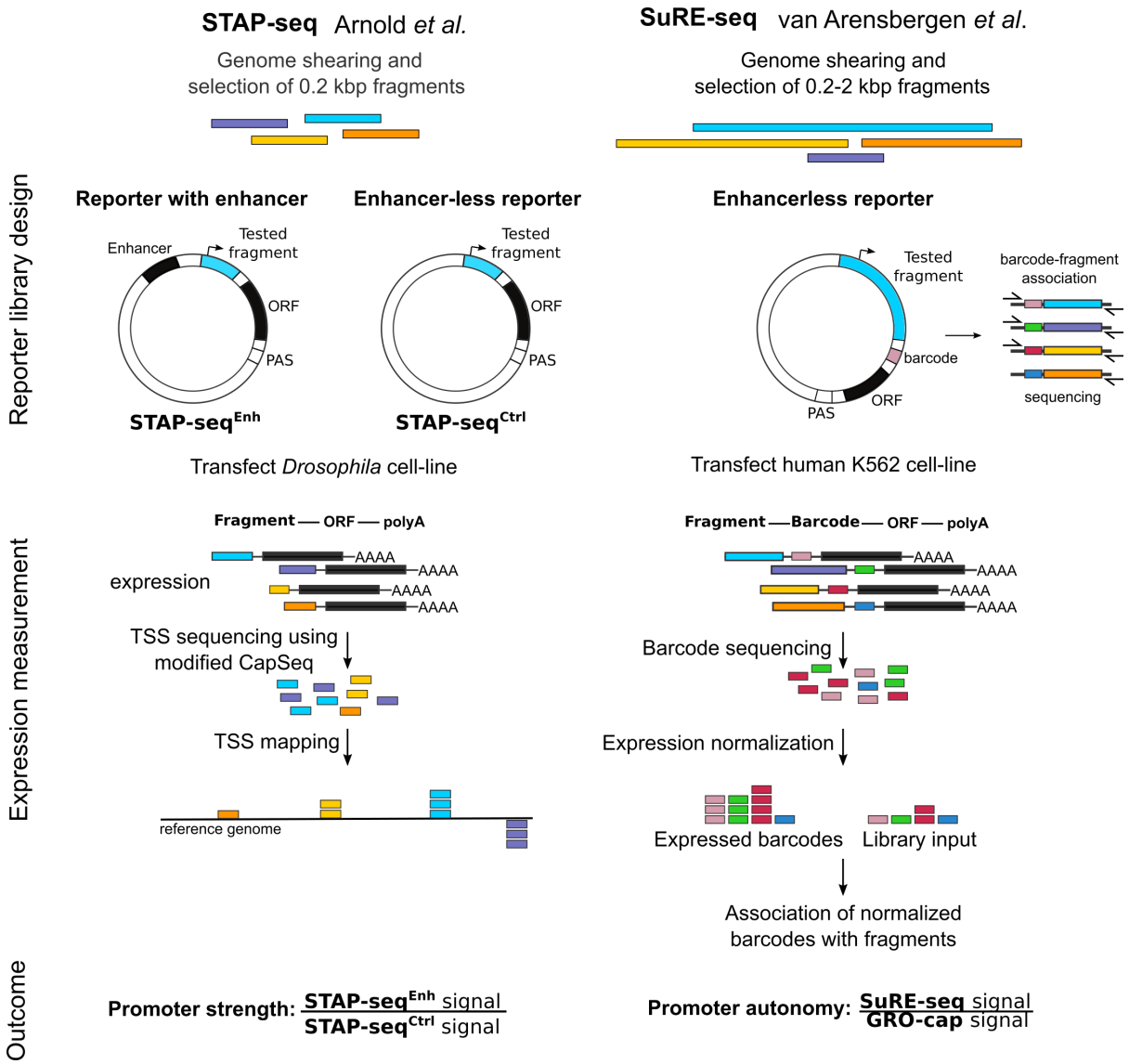


Figure 1. Comparison of strategies for measurement of promoter activity. In STAP-seq, *Drosophila* genome is sheared, 0.2kbp fragments selected and cloned into a reporter construct with or without an enhancer (STAP-seq^{enh} or STAP-seq^{ctrl}). SuRE-seq selects 0.2-2 kbp fragments and uses only an enhancer-less reporter. In STAP-seq, expression is measured by selection of capped mRNA and 5'-tag sequencing (CapSeq) which allows TSS mapping. SuRE methodology requires sequencing of the reporter library to associate barcodes with genomic fragments. Expression is measured by barcode sequencing and normalized per amount of library plasmids in transfected cells. The main readout of STAP-seq is promoter strength, defined as the ratio of expression obtain in enhancer *vs* enhancer-less reporter setup. In SuRE-seq, the main readout is promoter autonomy, defined as the ratio of autonomous (SuRE-seq signal) and endogenous expression (GRO-cap signal).