Europe PMC Funders Group
Author Manuscript
Nat Methods. Author manuscript; available in PMC 2012 May 01.

Published in final edited form as:
Nat Methods. ; 8(11): 963–968. doi:10.1038/nmeth.1705.

The proteomes of transcription factories containing RNA polymerases I, II or III

Svitlana Melnik[1,3], **Binwei Deng**[1], **Argyris Papantonis**[1], **Sabyasachi Baboo**[1], **Ian M. Carr**[2], and **Peter R Cook**[1]

[1]Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford, OX1 3RE, UK

[2]Leeds Institute of Molecular Medicine, University of Leeds, St. James's Hospital, Beckett Street, Leeds, LS9 7TF, UK

Abstract

Human nuclei contain three RNA polymerases (I, II, and III) that transcribe different groups of genes; the active forms of all three are difficult to isolate because they are bound to the substructure. Here, we describe a purification approach for isolating active RNA polymerase complexes from mammalian cells. After isolation, we analyzed their protein content by mass spectrometry. Each complex represents part of the core of a transcription factory; for example, the RNA polymerase II complex contains subunits unique to RNA polymerase II plus various transcription factors, but shares a number of ribonucleoproteins with the other polymerase complexes; it is also rich in polymerase II transcripts. We also describe a native chromosome conformation capture method to confirm that the complexes remain attached to the same pairs of DNA templates found *in vivo*.

INTRODUCTION

Eukaryotic nuclei contain three RNA polymerases (I, II, III) which are currently defined by the sets of genes they transcribe[1]. Polymerase I produces 45S rRNA (a precursor of 18S and 28S rRNA), polymerase II transcribes most genes encoding proteins, and polymerase III makes various small RNAs (including 7SK small nuclear RNA and tRNAs). The core of each polymerase has been purified and the structure determined, and we now have detailed knowledge of the way each works *in vitro*[2]. They are part of larger complexes; for example, the polymerase II complex is also involved in capping, splicing, and polyadenylation[3,4]. These mega-complexes may, in turn, be organized into larger "factories" that contain high concentrations of most machinery required for transcript production[5,6]. Transcription factories are defined as nuclear sites containing at least two different, active, transcription units[5]. However, the existence of such factories remains controversial, and one reason for this is that they have not yet been isolated[7].

Correspondence: Peter R Cook, Telephone: (+44/0) 1865 275528, Fax: (+44/0) 1865 275515, peter.cook@path.ox.ac.uk.
[3]Present Address: Division of Molecular Biology of the Cell II, German Cancer Research Center, DKFZ-ZMBH-Alliance, INF 581, D-69120, Heidelberg, Germany

AUTHOR CONTRIBUTIONS

Experiments were designed by SM, BD, AP, SB, and PRC. SM developed the isolation procedure and carried out many of the validation experiments, SM and BD performed gel electrophoreses and mass spectrometry, AP developed native 3C and carried out RT-PCR, SB did the light microscopy, and IMC developed software. All authors wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Note: Supplementary information is available on the Nature Methods website.

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Much of our knowledge concerning transcription was obtained using isolated polymerase cores assayed on exogenous templates. Two factors make purification of mammalian polymerases engaged on endogenous templates difficult. First, active enzymes represent a quarter of the total enzyme population; most are part of a rapidly-diffusing soluble pool that aggregates in non-isotonic buffers[8,9]. Therefore, we used isotonic conditions when removing the inactive fraction. Second, engaged polymerases plus their templates and transcripts are housed in factories bound to the underlying nuclear substructure[9,10]. Thus, a typical polymerase I factory in HeLa cells contains ~4 ribosomal cistrons transcribed on the surface of a "fibrillar center", embedded with others in a nucleolus[8]. Whole nucleoli can be freed from the substructure and purified, and mass spectrometry has yielded a detailed inventory of their contents[11]. Active polymerases II and III are found in dedicated nucleoplasmic factories, and polymerase II factories have been characterized in detail; high-resolution imaging[12] and quantitative analyses[8] have shown that one polymerase II factory typically contains ~8 polymerizing complexes on the surface of a polymorphic protein-rich core (average diameter ~90 nm, mass ~10 MDa). As caspases deconstruct nuclei during apoptosis, we reasoned they might be used to release factories from the substructure. [Core subunits of the three polymerases lack sites recognized by the caspases used, except RPB9.]

Here, we describe an approach for partial purification and characterization of the three transcription factory complexes from mammalian cells. All have apparent molecular masses of > 8 MDa, the size of the largest protein marker available. Each contains a characteristic proteome, as well as shared components. We also develop a method, native 3C (chromosome conformation capture), to validate that these complexes are not aggregation artifacts. With native 3C we show that isolated complexes remain associated with the same templates as found *in vivo* by conventional 3C.

RESULTS

Purification approach

To develop a method to purify transcription factories (Fig. 1a**)**, we begin by permeabilizing HeLa cells in a "physiological buffer" (PB); essentially all transcriptional activity is retained[8] as the inactive pool is lost[9]. Next we isolate nuclei using NP40, treat them with DNase I, and centrifuge the sample to leave most inactive chromatin in the supernatant. The pellet is next resuspended in "native lysis buffer" (NLB), treated with caspases to release large fragments of transcription factories, and respun (Supplementary Fig. 1 illustrates experiments used to optimize release). The supernatant is retreated with DNase to degrade residual chromatin.

As polymerase II activity is associated with a ~10-MDa core[12], we tested various techniques for purifying large complexes. Free-flow electrophoresis (both zone and isotachophoresis) failed to resolve different complexes. Sedimentation through sucrose or glycerol gradients allowed purification of a minority of polymerase I in polymorphic ~100-nm complexes (Supplementary Fig. 2), without resolving polymerase II and III complexes (which sediment less rapidly). Electrophoresis in "blue native gels"[13] was more successful. After running a second dimension without Coomassie blue, three partially-overlapping complexes were resolved; all ran slower than the largest (8 MDa) protein marker available.

Recovery of nascent RNA was monitored during purification by allowing polymerases in permeabilized cells to extend their transcripts by "running on" in [^{32}P]UTP by < 40 nucleotides[8]; then, ~85% of the resulting [^{32}P]RNA pellets after treatment with DNase I (in fraction "4pellet"; Fig. 1b). About half this (nascent) [^{32}P]RNA can be released by a set of caspases (into fraction "5super"; Fig. 1b). Significant amounts of run-on activity are also released, but determining how much is complicated by truncation of endogenous templates

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

by DNase I and transfer to NLB which halves run-on activity (in Fig. 1c, compare recoveries obtained after transfer to NLB). Nevertheless, 25% of the original activity remains in the "5super" fraction (Fig. 1c) – equivalent to ~50% after correction for losses due to the buffer. Immunoblotting confirmed that much of polymerases I and II was retained in "5super", whereas more polymerase III was lost (Supplementary Fig. 1d).

Polymerizing complexes of > 8 MDa

After 2D gel electrophoresis, complexes containing nascent [^{32}P]RNA and protein were found along the diagonal; immunoblots revealed that the three polymerases were partially resolved and ran as overlapping complexes of > 8 MDa (Fig. 2a). We named these complexes I, II, and III after the polymerases they contain. Complex I ran the fastest, even though it also sedimented the fastest in sucrose gradients (Supplementary Fig. 2). We traced this discrepancy to a destabilization induced by the Coomassie blue in the first dimension. In the absence of the stain, complex I runs the slowest (Fig. 2b), so we use Coomassie-free gels when purifying complex I. Excised regions of 2D gels enriched in the different complexes contained different proteins (Fig. 2c).

Proteomes of the complexes

We analyzed the protein content of the transcription factory complexes by liquid chromatography followed by tandem mass spectrometry. We identified peptides using a pipeline[14] that combines three search engines to provide a significantly lower false discovery rate (FDR); even so, we selected a conservative FDR of <1%. We detected several hundred proteins in each complex – some unique, others shared (Fig. 3a; Table 1, Supplementary Table 1).

Complexes I and II contained three and five subunits unique to RNA polymerases I and II, respectively (Table 1). Complex III contained one subunit shared by polymerases I and III (RPAC1), but none unique to polymerase III – consistent with the losses seen in fraction "3super" (Supplementary Fig. 1d). Each complex possessed a characteristic set of proteins (Table 1, Supplementary Table 1). Reassuringly, 83% proteins identified in complex I are also present in the proteome of isolated nucleoli[11]. Complex II contained general transcription factors like AP-2, CEBPB, and TFIIH (represented by ERCC3), specific regulators like CTCF and SAFB/B2, and histone methyltransferases (EZH2, SUV39H1, SUV39H2). Complex III contained Lupus La antigen (a polymerase III factor).

All three complexes share proteins involved in DNA/RNA metabolism including helicases, nucleic acid/nucleotide binding proteins, ribonucleoproteins (RNPs), and structural proteins like spectrin and actin (Table 1, Supplementary Table 1). Many are probably essential constituents of all complexes, others are likely cross-contaminants (for example, polymerase I/III specific subunits RPA2, RPA12 and RPAC1 in complex II) resulting from incomplete resolution in the gel.

As determining absolute amounts of proteins by mass spectrometry remains challenging, we used the normalized spectral index method to estimate relative abundancies[15]. Structural proteins were among the most abundant proteins (Supplementary Table 2), including RNA-binding proteins (the snoRNP dyskerin, hnRNPs H and K), spectrins and lamins in complex I, nucleophosmin in complex II, and alpha-actinin-1 in complex III.

Analysis of GO terms

More than half the proteins in each complex are associated with the gene ontology (GO) term "gene expression" (Fig. 3a,b), and each complex contained many proteins with expected terms. For example, complex II contained more proteins with "transcription from

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

RNA polymerase II promoter" (GO: 0006366) than the others (Fig. 3b). To place analysis on a more systematic basis, we compared GO terms associated with our proteins and the 87,130 terms in a database of all human proteins, or the 9,682 just associated with the GO term "nucleus" (Supplementary Fig. 3). We found that, for example, the 5 most over-represented terms for the transcription factory proteins compared with all human proteins had obvious connections with transcription, with terms "RNA binding", "RNP complex", and "RNA processing" heading the lists in the GO domains "molecular function", "cellular components", and "biological processes", respectively (Supplementary Fig. 3a). Compared to all human proteins, complex II also contained more terms associated with "gene expression" (GO: 0010467; 300 proteins; $P < 10^{-109}$; see Methods for the statistical test used), "transcription" (GO: 0006251; 149 proteins; $P < 10^{-54}$), "splicing" (GO: 0008380; 114 proteins; $P < 10^{-65}$), and "polyadenylation" (GO: 0043631; 3 proteins; $P < 10^{-3}$) – as well as processes closely coupled to (polymerase II) transcription like "DNA replication" (GO: 0006260; 58 proteins; $P < 10^{-19}$) and "DNA repair" (GO: 0006281; 76 proteins; $P < 10^{-24}$). Complex I was enriched in proteins with terms "ribosome biogenesis" (GO: 0042254; 88 proteins; $P < 10^{-98}$) and "rRNA processing" (GO: 0006364; 61 proteins; $P < 10^{-64}$).

To determine which GO terms concisely describe all proteins in the complexes, we developed a software tool, "MS-prot", which links UniProt accession numbers to associated GO terms. We combined selected terms (for example "mRNA cleavage", "splicing") into one user-defined group ("RNA processing"); almost all terms associated with our complexes can then be contained in only seven groups related to transcript production (the group "other terms" contains the remainder). Finally, we expressed the number of terms in each group as a fraction of terms in all groups (Fig. 3c, left); proteins in the database associated with terms like "cytoplasm" and "nucleus" serve as controls (Fig. 3c, right). Our complexes yielded different patterns from controls, there appear to be few contaminants (as "other terms" is small), and "RNA processing" is the largest. The "nucleolus/translation" group is also large; this was expected as active polymerases I/III are found in/on nucleoli where ribosomes are assembled[16], and nascent RNA made by polymerase II colocalizes with > 20 ribosomal proteins[17]. Taken together, this analysis suggests that each complex possesses a distinct set of proteins (relevant transcription, processing factors), a large pool of shared ones (RNPs), and few external contaminants.

Confirming selected associations

We next confirmed that some proteins seen by mass spectrometry co-immunoprecipitated with nascent RNA; polymerase II (a positive control), ribosomal protein RPS6, nonsense-mediated decay factor RENT1, and a protein found in many nuclear complexes (PCNA) all co-immunoprecipitated with nascent RNA (Supplementary Fig. 4a). We used immunofluorescence (applied conventionally, and coupled to proximity ligation and antibody blocking) to confirm that proteins found only in complex II (for example, CTCF, Sp3, ATRX) were found in close proximity to active RNA polymerase II, others only in complex III (Lupus La, EXOSC6) lay close to polymerase III (although some Lupus La was found near polymerase II), and still others in all three complexes (DDX1, hnRNP A2/B1, U2AF65) lay close to both polymerases II and III (Supplementary Fig. 4b,c).

We also examined whether each complex contained the expected nascent RNAs using quantitative reverse-transcriptase PCR and intronic probes; for example, complex I contained ≥ 33-fold more nascent 45S rRNA than the other complexes (Supplementary Fig. 5a). The different complexes were also still associated with expected DNA fragments (inevitably some DNA survives DNase I treatment). Complex I contained relatively more DNA encoding 45S rRNA than the other two, complex II was richest in 3 genes transcribed

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

by polymerase II (*RPS6, ARHGAP5, MIR191*), and complex III contained the highest amounts of two polymerase III genes (*RN7SK, tRNA*-leuCAA; Supplementary Fig. 5b).

"Native 3C": the structure in complex II is like that *in vivo*

Our purification strategy (Fig. 1a**)** yields largely template-free complexes. However, treatment with *Hind*III (instead of DNase I) enables complexes containing more DNA to be isolated – albeit at the cost that the three complexes can no longer be resolved (Supplementary Fig. 6a). We therefore developed a new method to show that complexes are associated with the same active templates found *in vivo*.

Chromosome conformation capture (3C) is a powerful tool for detecting proximity of two DNA sequences in 3D space[18]; it involves fixation to cross-link DNA sequences lying together (Fig. 4a). In "native 3C" (Fig. 4a), we omit fixation, and rely on the natural interactions holding sequences together[19]. Here, we treat nuclei with *Hind*III to remove most DNA, released the complexes with caspases, ran the gel (which separates inactive DNA fragments from transcribed fragments attached to complexes), excised the relevant region (which now contains a diluted solution of factories and associated DNA embedded in agarose), added ligase to the gel, recovered the DNA, and detected novel ligation products by PCR.

For this experiment we used human umbilical vein endothelial cells (HUVECs) because we previously analyzed (by 3C) the changing contacts between a number of their genes induced by tumor necrosis factor alpha (TNFα)[20]. *NFKB1A, SAMD4A, TNFAIP2*, and *PTRF* are normally silent in HUVECs, but 30 min after adding TNFα they become active. Then, the 5′ end of *SAMD4A* comes to lie near *TNFAIP2* (on the same chromosome) and *PTRF* (on a different chromosome)[20]. We first confirmed these 3C results. Before adding TNFα, interactions 1-6 illustrated in Figure 4b did not yield bands on a gel (Fig. 4c). But after 30 min, interactions 1 and 5 – where both partners are responsive genes – yielded bands indicative of contacts (Fig. 4c). Interaction 2 remained undetected; we previously showed that this is because 221-kbp *SAMD4A* is so long that the first polymerase to begin transcribing it after stimulation does not reach the region involved in interaction 2 until ~85 min after stimulation – and only then are contacts with *PTRF* or *TNFAIP2* detected[20]. Interaction 3 (involving a constitutively-active gene lying immediately next to responsive *SAMD4A*), interaction 4 (involving two responsive genes lying 20 Mbp apart on the same chromosome), and interaction 6 (involving an as-yet untranscribed part of *SAMD4A* and another responsive gene) also remain undetected (Fig. 4c). These results confirm those obtained earlier[20], and are consistent with some TNFα-responsive genes (but not others), and some parts of responsive genes (but not others), coming together to be transcribed in the same dedicated factory[20].

Native 3C yields exactly the same pattern as 3C (Fig. 4c). Therefore, we conclude the contacts detected in isolated complexes are the same as those *in vivo* and so are unlikely to result from artifactual aggregation. Moreover, these interactions are specific, as 3C and native 3C yield no bands using primers targeting: (i) two responding but non-associating genes (Fig. 4c, interaction 4) – so contacts do not result simply from an aggregation of active genes, (ii) a polymerase II gene (*PTRF*) and either the (repeated) polymerase I rDNA gene or a polymerase III gene (*RN7SK*) – so contacts do not result simply from the effects of high copy number or hyper-activity, and (iii) the polymerase I gene (rDNA) and a polymerase III gene (*RN7SK*; Supplementary Fig. 6b) – so contacts again do not result from the effects of high copy number or hyper-activity. Significantly, less DNA prepared by native 3C gives bands of equivalent intensity (Fig. 4c, compare loadings for interactions 1 and 5) – consistent with fragments still attached to factories being purified away from unattached ones (Fig. 4a).

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

These results also show our general purification strategy can be extended to a different cell type (i.e., HUVECs). Finally, we exploit our ability to switch on transcription of selected genes in HUVECs to confirm that (residual) relevant templates are only found in complex II when transcribed. Thus, when uninduced, *SAMD4A, EXT1*, and *MIR17* are inactive[20] and not found in complex II; however, when induced by TNFα, they are enriched in complex II (but not complex III; Supplementary Fig. 6c).

DISCUSSION

The existence of transcription factories has been controversial, and one reason given for this is that they have not yet been isolated[7]. Here we have reported a method to isolate and analyze the proteomes of transcription factories. We suggest these complexes represent large fragments of factory cores still bound to the substructure. We anticipate that individual complexes in the pool we call complex II will be heterogeneous, as different types of nucleoplasmic factories are being uncovered[5,6].

In vitro systems for transcribing mammalian genes remain inefficient; the efficiency of our system could be increased by adding purified factors and endogenous templates to our complexes. However, two major difficulties remain. First, we have been unable to recover complexes from 2D gels without aggregation. Second, added templates will also have to displace tightly-bound endogenous ones. As a result, recovered “complexes” only have the usual low transcriptional activity on added templates.

Native 3C may prove to be a useful alternative to 3C for various applications (Fig. 4a). It mainly detects contacts between active alleles in the population – which may be the minority[6,20] – as most inactive alleles are lost during isolation. Background in native 3C may also be lower, as chemical fixation can stabilize adventitious contacts (Fig. 4a**)**, much of the DNA distant from (contact-rich) nodes is discarded during isolation, and less template is required for detection (Fig. 4c).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank J Bartlett (for technical assistance), M Vigneron (for antibodies), B Thomas, D Trudgian, G Ridlova, and M Dreger (for help with proteomics), M Shaw (for help with electron microscopy), and the Medical Research Council (SM, BD), EP Abraham Research Fund (BD), Biotechnology and Biological Sciences Research Council (AP), Wellcome Trust (AP), and Felix Scholarship Trust of Oxford University (SB) for support.

APPENDIX

METHODS

Cells, general procedures

Monolayer cells were grown in DMEM (Invitrogen) + 5% fetal calf serum (FCS; Biosera); suspension HeLa cells were grown in S-MEM (Invitrogen), 5% FCS, non-essential amino acids, 2 mM L-glutamine, and 11 mg/ml sodium pyruvate (all from PAA Laboratories Gmbh). HUVECs from pooled donors (Lonza) were grown to 80-90% confluency in Endothelial Basal Medium 2-MV with supplements (EBM; Lonza). Recoveries of DNA were measured by scintillation counting after growing cells in [methyl-^{3}H]thymidine (0.25 μCi/ml; ~50 Ci/mmol) overnight[10]. Unless stated otherwise, all buffers used with permeabilized cells were treated with diethylpyrocarbonate (DEPC) or prepared with DEPC-

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

treated water and maintained ice-cold, and all washes/spins were at 400 x g for 5 min at 4°C. The amount of protein in the area of a gel containing three complexes (Fig. 2a,ii – dotted oval) was measured by densitometry using AIDA software and blue carrier immunogenic protein (8 MDa; Pierce) as a standard. Recoveries of [^{3}H]DNA and [^{32}P]RNA in the same areas were measured by scintillation counting. Protein concentrations were monitored using a Nanodrop ND-1000 spectrophotometer (LabTech).

Permeabilization, "run-on" in [^{32}P]UTP

Run-on transcription was performed using triphosphate concentrations limiting elongation to < 40 nucleotides[8]. In brief, HeLa cells were permeabilized with saponin (170 μg/ml; 5 min; Sigma) in "physiological buffer" (PB). PB (pH 7.4) contains 100 mM potassium acetate, 30 mM KCl, 10 mM Na_2HPO_4, 1 mM $MgCl_2$, 1 mM Na_2ATP, 1 mM dithiothreitol, 25 units/ml RNaseOUT (Invitrogen), 10 mM β-glycerophosphate, 10 mM NaF, 0.2 mM Na_3VO_4, and a 1/1000 dilution of protease inhibitor cocktail (PIC; Sigma). As the acidity of ATP batches varies, 100 mM KH_2PO_4 was used to adjust the pH. After pelleting, the supernatant is called fraction "2super". Permeabilized cells in the pellet were now respuspended in PB, incubated (5 min on ice), and pelleted; this process was repeated three times. After resuspension again in PB, permeabilized cells were pre-incubated (33°C; 3 min), and a run-on performed (10 min; 33°C) in 100 μM ATP, 100 μM CTP, 100 μM GTP, 0.1 μM UTP, 50 μCi/ml [^{32}P]UTP (3,000 Ci/mmol; Perkin Elmer), and $MgCl_2$ giving a concentration of Mg^{2+} ions equimolar to triphosphates. Reactions were stopped by transfer to ice and immediate addition of EDTA to 2.5 mM. Incorporation of ^{32}P into acid-insoluble material, and subsequent recoveries of [^{32}P]RNA (as in Fig. 1b), were measured by scintillation counting[10]. Permeabilized cells were washed twice with PB to remove unincorporated label before factories were isolated.

Isolating factories

Caspases release polymerases bound to the nuclear substructure more efficiently from HeLa growing in suspension compared to monolayers, so suspension HeLa were used unless stated otherwise. Cells were permeabilized with saponin, and washed 4 times in PB; in some cases, a run-on in [^{32}P]UTP was performed and cells washed twice to remove free label (as above). After resuspension, permeabilized cells were lysed (5 min) in PB plus 0.4% NP40, and spun; the supernatant is called fraction "3super". Nuclei in the pellet were washed twice in PB + NP40 (with 5-min incubation on ice after each resuspension, as above) to give "3pellet". Resuspended nuclei were digested (30 min; 33°C) with either (i) DNase I (protease- and RNase-free; Worthington;10 units/10^7 cells in 100 μl PB plus 0.5 mM $CaCl_2$), or (ii) *Hae*III (Invitrogen; 1000 units/10^7 cells), or (iii) *Hind*III (New England Biolabs; 1000 units/10^7 cells) in PB; reactions were stopped by adding EDTA to 2.5 mM and cooling in iced water. Chromatin-depleted nuclei were spun (600 g; 5 min), and the supernatant ("4super") collected. The pellet ("4pellet") was resuspended (10^7 cells/100 μl) in "native lysis buffer" (pH 7.4; NLB). NLB was modified from Novakova *et al.*[13] and contained 40 mM Tris-acetate, 2 M 6-aminocaproic acid (Fluka), 7% sucrose, 1/1000 dilution of PIC, and 50 units/ml RNaseOUT. After 20 min, recombinant caspases 6, 8, 9 and 10 (Calbiochem or Biovision; a total of 2 units in NLB per 10^7 nuclei) were added; after 30 min at 33°C, the reaction was stopped with caspase inhibitor III (0.2 mM; Calbiochem), the solution spun (600 g, 5 min), and the supernatant ("5super") and pellet ("5pellet") collected. "5super" was now treated with DNase I (as above), EDTA (to 2.5 mM; the sample was now split into aliquots, frozen rapidly in dry ice, and stored at −80°C. Conditions for electrophoresis in a native 2D gel were modified from those used previously[13,21] by increasing the pore size of the gel, modifying the running buffer (to retain run-on activity), and reducing the concentration of Coomassie blue used to provide charge to the hydrophobic complexes analyzed originally. Composite (analytical) gels contained 1.5% acrylamide and

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

0.7% agarose (SeaKem Gold, Lonza) in 40 mM Tris-acetate (pH 7.4), 7% sucrose, and 0.01% Triton X-100, and were run (~1 h; 100 V; constant voltage) in 40 mM Tris-acetate (pH 7.4). A sample with bromophenol blue and xylene cyanol (both added to 0.04%) was run until the xylene cyanol reached ¾ of the length (and bromophenol blue is lost). For the "blue" version, 0.02% and 0.002% Coomassie blue G-250 were added to sample and cathode buffers used in the first dimension, respectively. After running the first dimension, the lane containing the sample was cut out of the gel and polymerized with the second dimension using the same gel and buffers as in the first. For preparative gels used for mass spectrometry, "5super" (from $5\times^{7}$ cells unlabeled with ^{32}P) was applied to a gel lacking Triton X-100, runs (overnight; 4°C) began with 100 V (until the sample entered the gel) and then at 40 V. Blue carrier immunogenic protein (8 MDa; Pierce) was used as a marker. Gels were stained with Coomassie blue (Imperial protein stain, Pierce).

Mass spectrometry

After fractionation on 2D gels, regions corresponding to those rich in [^{32}P]RNA and one of the polymerases (detected by autoradiography and immunoblotting using analytical gels run in parallel) were excised, equilibrated (10 min) in 2 changes of 1x Tris-glycine running buffer, loaded on a SDS-acrylamide gel, and subjected briefly to electrophoresis so that all denatured proteins just enter the resolving gel. The whole sample was excised as one gel piece, treated with trypsin, the resulting peptides extracted, vacuum dried, and injected (usually 3 injections/sample; 120 min/injection) into a Dionex U3000 nanoHPLC system coupled to a Thermo LTQ Orbitrap mass spectrometer. The three resulting raw data files were merged, converted to .mzXML format using ReAdW v4.2.1 (http://tools.proteomecenter.org/wiki/index.php?title=Software:ReAdW), and submitted to the Central Proteomics Facilities Pipeline[14] (CPFP). Mass spectrometry data is typically analyzed using a single search engine like "Mascot" (Matrix Science). CPFP uses multiple search engines, modeling tools, and target-decoy validation to provide peptide/protein identifications with significantly higher confidence; this provides a stringent test, and proteins in complexes I, II, and III were identified with false discovery rates (FDRs) below 1%. Briefly, .mzXML files are submitted to "Mascot", "X! Tandem" (used with the k-score plugin[22], and the Open Search Algorithm[23]; resulting peptide identifications were then validated with "PeptideProphet"[24]. "iProphet" was used to combine peptide "hits" from the three search engines and to refine identification probabilities according to additional criteria[25]. All searches were performed against a concatenated target/decoy database (International Protein Index human v3.64; precursor mass tolerance ± 10 ppm; fragment mass tolerance ± 0.5 Da; fixed modification – carbamidomethyl for C; variable modifications – acetylated protein for N-terminal, deamidation for N and Q, and oxidation for M), providing empirical FDRs[26] that were compared with estimated ones from the "Prophet" tools to validate results. By default, results are reported at a 1% target/decoy FDR for both peptides and proteins. For results shown in experiment 1, 90%, 95% and 97% proteins in complexes I, II, and III respectively were retained when the FDR filter was set more stringently at 0.5%. Two additional experiments (experiments 2 and 3) were also conducted; in both, blotting showed polymerases were less well resolved, and in experiment 3 complex I was not analyzed. 73%, 60%, and 81% proteins seen in the first experiment in complexes I, II and III, respectively, were also seen in the second. 39% and 53% proteins seen in the first experiment in complexes II and III, respectively, were also seen in the third (in which fewer proteins were seen). Details of the contents of each complex can be found in Supplementary Tables 4,5,6,7, and complete proteomic datasets are available at http://users.path.ox.ac.uk/~pcook/data/ContentOfFactories.html and https://proteomecommons.org/tranche/ using the following hash codes: 'read me' file, lysDE6I7cXJA140DP5FCpSYtJKPBWgUUNmOgyTBb04HNd7DKVVzzbzWcUCgho9lrypjaIQWMnN0Zfg0Z+WN0fJk1mc8AAAAAAAABYw==, (ii) experiment 1,

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

IqeHRUGUiEPR4v7WLY0epG4aSLRYid4aCBkJ6ZHYpxzoxb89gRcrX+RQ/98alnP7VT4DVAQLnRLvMW902MsqHyzn5fYAAAAAAAAAZpg==, (iii) experiment 2, v3Wi7PA3krKsjlqA241eRfMWMcyu8pYnqIimft82ZnZLm39F0BfrmYc/Aguo8jYMR6u1sU8z+rDGx4adsF4BjgqblDYAAAAAAAAM0w==, (iv) experiment 3, pAF+fdNbP/2tkcWx1huqyHhoUejqQTera1UfRnDSHIIPhFPrjDn8V7eu7+fA8PGJ3F1GZXSYLU7RYYoBjLplwJRoVTEAAAAAAAARuA= and (v) comparing complexes I, II and III (seen in all three experiments), I7Cdw8venrUMm8VWOsg5H0sKzCd58MdiJ+n3+Hn3PM1BS6It5NypoQKFNiTGIiRSjNr4xNc32woycFb4Q8TNpB99+HgAAAAAAAC+w==.

GO term analysis

To analyze complex content, protein identifications were exported from CPFP into "ProteinCenter" (Proxeon); FDR filters of 0.82%, 0.8%, and 0.84% (average FDRs of each dataset) were maintained throughout analysis in "ProteinCenter". Over- and under-represented GO categories (Supplementary Fig. 3) were obtained by comparison of frequencies seen with those obtained with either a standard set of all human proteins (i.e., the > 87,000 entries in the human International Protein Index; http://www.ebi.ac.uk/IPI/IPIhelp.html) or the 9,682 (nuclear) proteins obtained by filtering this database with the GO term "nucleus" (GO: 0005634). *P* values relating to significance of any differences seen were evaluated using the statistical test incorporated into "ProteinCenter"[27]. To compare GO terms associated with complexes (Fig. 3c), we developed software ("MS-prot"; http://www.ms-prot.co.uk; available free) that connects an UniProt accession number in a protein database to associated GO terms, and allows the user to define a group of GO terms and filter out proteins linked to terms in the group. The group "Transcription" contained GO terms "RNA polymerase", "transcription factor", and "transcription regulation"; group "RNA processing" contained terms "exosome", "mRNA cleavage", "mRNA polyadenylation", "nonsense mediated decay", "RNA binding", "RNA helicase", "RNA metabolism", "RNA modification", and "splicing"; group "RNPs" contained term "ribonucleoprotein"; group "DNA/chromatin" contained terms "DNA binding", "DNA topology", "DNA helicase", "DNA replication", "DNA damage", and "DNA repair"; group "nucleolus/translation" contained terms "nucleolus", "ribosome", "ribosome biogenesis", and "translation"; group "nucleotide binding" contained terms "nucleotide binding" and "nucleoside binding"; group "kinases/phosphatases" contained terms "kinase" and "phosphatase"; "other terms" contained all those not included above. Four other sets of proteins are included for comparison: (i) 18,679 proteins associated with the term "cytoplasm"(GO: 0005737), and 9,682 proteins with the term "nucleus" (GO: 0005634) from the International Protein Index (above), (ii) 4,666 proteins from the nucleolus database[11] http://www.lamondlab.com/NOPdb3.0/), and (iii) 67 "S100" proteins obtained by filtering entries in the Uniprot database (http://www.uniprot.org/) with the key word "S100".

Protein quantification

Label-free relative quantification of proteins within samples was performed using the normalized spectral index (SI) method[15] which combines three abundance features (peptide count, spectral count, fragment-ion intensity). SIs were calculated using the output from one search engine – "Mascot" – using the default significance setting of < 0.05 and a script available on request. Use of a single search engine (not three as above) results in a slightly different list of proteins to that obtained with CPFP. To increase stringency, we selected proteins yielding ≥ 3 peptides; 89, 95 and 95% of the total SI in the output was retained at this stage for complexes I, II, and III, respectively. We then ranked surviving proteins according to their SI, and the top ten are listed in Supplementary Table 2. As these constitute 66%, 60% and 64% of the total SI seen in complexes I-III, respectively, we are confident

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

these 10 proteins are amongst the most abundant. The same top ten proteins were seen in complexes II and III in experiments 1 and 3 (SI analysis was not performed in experiment 2).

Native 3C

After an initial treatment with *Hind*III, the region of a gel containing complexes with more DNA (see Supplementary Fig. 6e) was excised, diced, incubated (4°C; 3 d) in ligation buffer (NEB), 1 mM ATP, T4 DNA ligase (2,000 units/ml; NEB), and DNA isolated using a MicroElute gel extraction kit (Omega Bio-Tek). 3C was then performed as described, using sets of validated primers targeting *SAMD4A* and *PTRF*[20]. Other primers were selected using Primer 3.0 (http://frodo.wi.mit.edu/primer3/) to have an optimal length of 20-22 nucleotides, a Tm of 62°C, and to yield amplimers of 100-200 bp (Supplementary Table 3). PCRs (25 μl reactions) were performed using GoTaq polymerase (Promega) with one cycle at 95°C for 2 min, followed by 36 cycles at 95°C for 45 s, 59°C for 45 s, 72°C for 20 s, and a final step of 72°C for 2 min. Amplimers were separated in 2.5% agarose gels, stained with SYBR Green, and scanned in an FLA-5000 scanner (Fuji). The hybrid nature of 3C/native 3C bands was verified by sequencing.

REFERENCES

1. Roeder RG. The eukaryotic transcriptional machinery: complexities and mechanisms unforeseen. Nat. Med. 2003; 9:1239–1244. [PubMed: 14520363]
2. Cramer P, et al. Structure of eukaryotic RNA polymerases. Annu. Rev. Biophys. 2008; 37:337–352. [PubMed: 18573085]
3. Das R, et al. SR proteins function in coupling RNAP II transcription to pre-mRNA splicing. Mol. Cell. 2007; 26:867–881. [PubMed: 17588520]
4. Shi Y, et al. Molecular architecture of the human pre-mRNA 3′ processing complex. Mol. Cell. 2009; 33:365–376. [PubMed: 19217410]
5. Cook PR. A model for all genomes; the role of transcription factories. J. Mol. Biol. 2010; 395:1–10. [PubMed: 19852969]
6. Chakalova L, Fraser P. Organization of transcription. Cold Spring Harb Perspect Biol. 2010; 2:a000729. [PubMed: 20668006]
7. Sutherland H, Bickmore WA. Transcription factories: gene expression in unions? Nat. Rev. Genet. 2009; 10:457–466. [PubMed: 19506577]
8. Jackson DA, Iborra FJ, Manders EMM, Cook PR. Numbers and organization of RNA polymerases, nascent transcripts and transcription units in HeLa nuclei. Mol. Biol. Cell. 1998; 9:1523–1536. [PubMed: 9614191]
9. Kimura H, Tao Y, Roeder RG, Cook PR. Quantitation of RNA polymerase II and its transcription factors in an HeLa cell: little soluble holoenzyme but significant amounts of polymerases attached to the nuclear substructure. Mol. Cell. Biol. 1999; 19:5383–5392. [PubMed: 10409729]
10. Jackson DA, Cook PR. Transcription occurs at a nucleoskeleton. EMBO J. 1985; 4:919–925. [PubMed: 2990913]
11. Ahmad Y, Boisvert FM, Gregor P, Cobley A, Lamond AI. NOPdb: Nucleolar Proteome Database--2008 update. Nucleic Acids Res. 2009; 37:D181–184. [PubMed: 18984612]
12. Eskiw CH, Rapp A, Carter DRF, Cook PR. RNA polymerase II activity is located on the surface of ~87 nm protein-rich transcription factories. J. Cell Sci. 2008; 121:1999–2007. [PubMed: 18495842]
13. Novakova Z, Man P, Novak P, Hozak P, Hodny Z. Separation of nuclear protein complexes by blue native polyacrylamide gel electrophoresis. Electrophoresis. 2006; 2:1277–1287. [PubMed: 16502463]
14. Trudgian DC, et al. CPFP - The Oxford Central Proteomics Facility Pipeline. Clinical Proteomics. 2009; 5(Supplement 1):94.

Europe PMC Funders Author Manuscripts

15. Griffin NM, et al. Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. Nat. Biotechnol. 2010; 28:83–89. [PubMed: 20010810]
16. Hopper AK, Pai DA, Engelke DR. Cellular dynamics of tRNAs and their genes. FEBS Lett. 2010; 584:310–317. [PubMed: 19931532]
17. Iborra FJ, Escargueil AE, Kwek KY, Akoulitchev A, Cook PR. Molecular cross-talk between the transcription, translation, and nonsense-mediated decay machineries. J. Cell Sci. 2004; 117:899–906. [PubMed: 14762111]
18. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. Science. 2002; 295:1306–1311. [PubMed: 11847345]
19. Cullen KE, Kladde MP, Seyfred MA. Interaction between transcription regulatory regions of prolactin chromatin. Science. 1993; 261:203–206. [PubMed: 8327891]
20. Papantonis A, et al. Active RNA polymerases: mobile or immobile molecular machines? PLoS Biol. 2010; 8:e1000419. [PubMed: 20644712]
21. Nadano D, Aoki C, Yoshinaka T, Irie S, Sato TA. Electrophoretic characterization of ribosomal subunits and protein in apoptosis: specific downregulation os S11 in staurosporine-treated human breast carcinoma cells. Biochemistry. 2001; 40:15184–15193. [PubMed: 11735401]
22. Maclean B, Eng JK, Beavis RC, McIntosh M. General framework for developing and evaluating database scoring algorithms using the TANDEM search engine. Bioinformatics. 2006; 22:2830–2832. [PubMed: 16877754]
23. Geer LY, et al. Open mass spectrometry search algorithm. J. Proteome Res. 2004; 3:958–964. [PubMed: 15473683]
24. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal. Chem. 2002; 74:5383–5392. [PubMed: 12403597]
25. Shteynberg, D., et al. iProphet: Improved Validation of Peptide and Protein Ids in the Trans-Proteomic Pipeline; Poster session at: HUPO 7th Annual World Congress; Amsterdam. August 16-20; 2008.
26. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods. 2007; 4:207–214. [PubMed: 17327847]
27. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Statist. Soc. Ser. B. 1995; 57:289–300.
28. Zheng B, Han M, Bernier M, Wen JK. Nuclear actin and actin-binding proteins in the regulation of transcription and gene expression. FEBS J. 2009; 276:2669–2685. [PubMed: 19459931]
29. Hou C, Corces VG. Nups take leave of the nuclear envelope to regulate transcription. Cell. 2010; 140:306–308. [PubMed: 20144754]

Europe PMC Funders Author Manuscripts

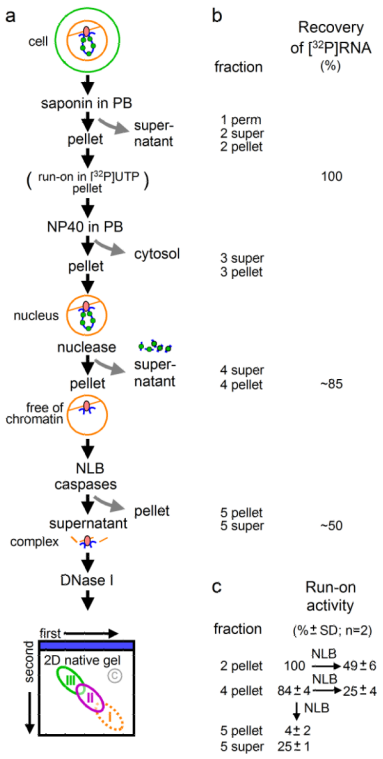

Figure 1.

Purification procedure.

(**a**) Strategy. Cartoon (top left): chromatin loop with nucleosomes (green circles) tethered to a polymerizing complex (oval) attached to the substructure (brown). Cells are permeabilized, in some cases a run-on performed in [^{32}P]UTP so nascent RNA can be tracked, nuclei are washed with NP40, most chromatin detached with a nuclease (here, DNase I), chromatin-depleted nuclei resuspended in NLB, and polymerizing complexes released from the substructure with caspases. After pelleting, chromatin associated with polymerizing complexes in the supernatant is degraded with DNase I, and complexes partially resolved in 2D gels (using “blue native” and “native” gels in the first and second dimensions); rough positions of complexes (and a control region, c) are shown. Finally, different regions are excised, and their content analyzed by mass spectrometry.

(**b**) Recovery of [^{32}P]RNA, after including a “run-on”. Fractions correspond to those at the same level in (**a**).

(**c**) “Run-on” activity assayed later during fractionation (as in **a**, but without run-on at beginning). Different fractions, with names as in (**a**), were allowed to extend transcripts by < 40 nucleotides in [^{32}P]UTP, and the amount of [^{32}P]RNA/cell determined by scintillation counting. Fractions “2pellet” and “4pellet” were also resuspended in NLB before run-ons were performed; results indicate NLB reduces incorporation to a half or less (right). Despite this, “5super” possesses 25% run-on activity of permeabilized cells (“2pellet”) – equivalent to half the original (after correction for effects of NLB).

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

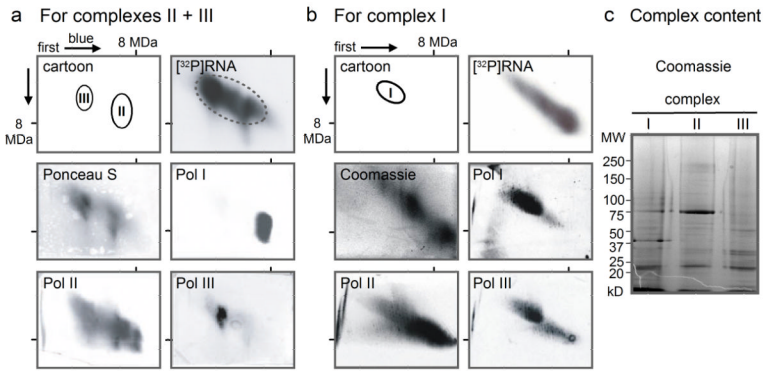

Figure 2.

Resolving different polymerases in "native" 2D gels (run-ons in [^{32}P]UTP included).

(**a)** Resolving complexes II + III with Coomassie blue in the first dimension. The cartoon shows regions selected for mass spectrometry analysis. First, an autoradiograph of the gel was prepared; overlapping spots of (nascent) [^{32}P]RNA are along the diagonal. ~0.03% protein, ~0.8% DNA, and ~5% nascent [^{32}P]RNA initially present were contained in the region indicated (dotted outline). After blotting, the membrane was stained with Ponceau S; most protein is on the diagonal. Next, the membrane was immuno-probed successively for three polymerases (using antibodies against RPA194, RPB1, and RPC62); the three are partially resolved. Note that complex I is destabilized by the Coomassie blue in the first dimension, and so migrates rapidly.

(**b**) Resolving complex I (no Coomassie in either dimension). The cartoon shows regions selected for mass spectrometry analysis. First, an autoradiograph was prepared; overlapping spots of (nascent) [^{32}P]RNA are again along the diagonal. After staining with Coomassie, spots are seen to overlap regions rich in [^{32}P]RNA. After blotting, the membrane was probed for the polymerases (as above); complex I now runs the slowest.

(**c**) Proteins in regions indicated in **a** and **b** were resolved on a 4-15% SDS-acrylamide gel, and stained with Coomassie.

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

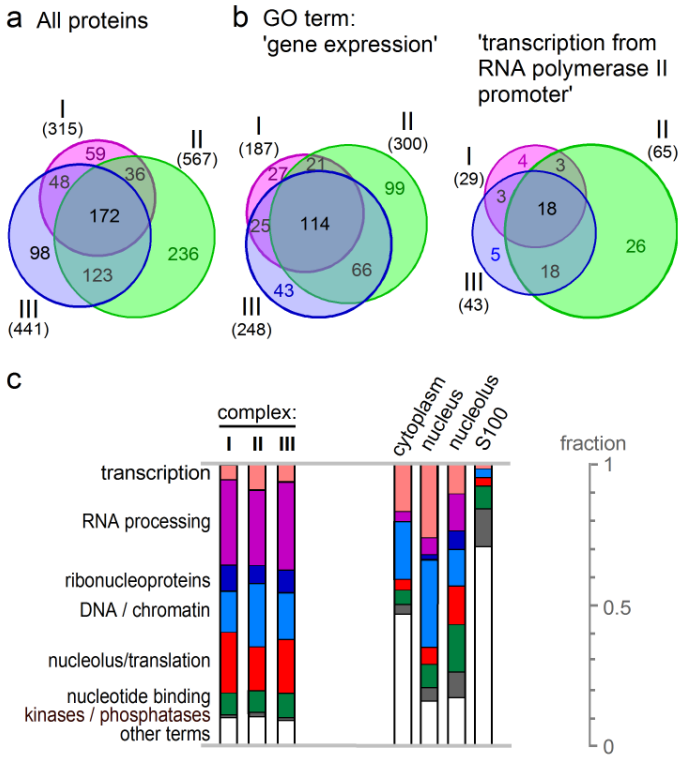

Figure 3.

The content of complexes I, II, and III determined by mass spectrometry.

(**a)** Numbers of proteins in the different complexes and their overlap.

(**b**) Many proteins in each complex are associated with the GO term “gene expression” (GO: 0010467), and complex II contains more with “transcription from RNA polymerase II promoter” (GO: 0006366) than I and III.

(**c**) Most proteins in each complex possess GO terms related to transcript production. Selected GO terms were incorporated into 8 groups; for example, “transcription” includes terms “RNA polymerase”, “transcription factor” and “transcription regulation”), and “other terms” includes those not in the other 7 groups. Four additional sets of proteins are included for comparison on the right. Some proteins possess terms in more than one group, and terms in each group are expressed as a fraction of the total in all groups. 2% proteins in each complex lacked any GO term, and many proteins in the complexes associated with “other terms” nevertheless turn out to play a role in transcript production (for example, actin[28], proteasomal constituents[17], nucleoporins[29]). Each complex exhibits a characteristic pattern, which is distinct from those given by proteins with the terms “cytoplasm” and “S100”.

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

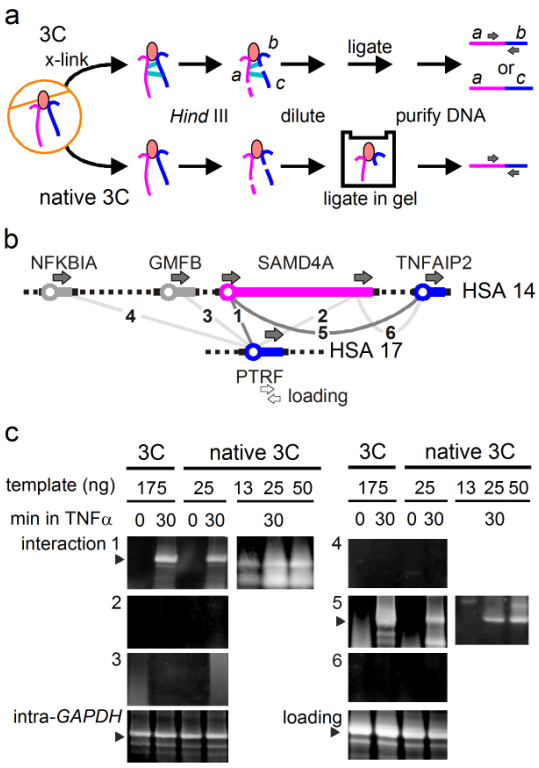

Figure 4. Isolated complexes remain associated with DNA sequences found *in vivo*

(**a**) Strategies for 3C and native 3C. Magenta and blue genes on different chromosomes are co-transcribed by one complex (oval) attached to the substructure (brown). 3C involves covalently cross-linking (turquoise lines) DNA, cutting (with *Hind*III here), dilution, ligation, and detection of ligated products by PCR. Note that *a* is joined to *c*, even though there was no stable molecular bridge between the two before cross-linking; such products yield an inevitable background. Native 3C omits cross-linking, and relies on pre-existing (native) contacts. As most (inactive) cellular DNA is lost during isolation (including fragment *c*), unwanted background is lower, and wanted 3C products are present in higher concentrations.

(**b**) Targets of primers (grey arrows) used to monitor interactions 1-6; only contacts due to interactions 1 and 6 (purple lines) are detected by both 3C and native 3C. White arrows: primers used for loading controls.

(**c**) 3C and native 3C yield similar bands/contacts (although less template is needed with native 3C). HUVECs were treated with TNFα (0, 30 min), and interactions 1-6 monitored by 3C and native 3C. Arrowheads indicate relevant 3C bands (all verified by sequencing; additional, non-specific, bands are amplified during the 36 PCR cycles used). “Intra-*GAPDH*” 3C and “loading” controls apply to all panels. Controls (with 13-50 ng template) show PCR is conducted in the linear amplification range.

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Table 1

A selection of proteins detected by mass spectrometry in the three complexes.

Complex (protein group)	Protein
Complex I	
RNA polymerase	RPA2; RPA34; RPA49; RPABC1.
Transcription regulators	LYRIC; ILF2; SMARCA4.
Complex II	
RNA polymerase	RPB2; RPB3; RPB4; RPB7; RPB9; RPABC3; RPA2[a]; RPA12 [a]; RPAC1 [a].
Transcription factors	Activator of basal transcription 1; TFII-I; TFIIH subunit 1; XPB helicase; TF20; TF AP-2 alpha; TF AP-4; TF Sp3; CCAAT/enhancer-binding protein-beta; CTCF; ATRX; USF1.
Transcription regulators	Scaffold attachment factors B1 and B2; SAFB-like transcription modulator; sex comb on midleg-like protein; splicing factor 1; SWI/SNF-related matrix-associated actin-dependent regulator; major centromere autoantigen B; far upstream element-binding protein 1; HMG20A; chromatin assembly factor 1 subunit B.
Histone modification enzymes	Histone-lysine N-methyltransferases EZH2, SUV39H1, and SUV39H2.
Complex III	
RNA polymerase	RPAC1.
Transcription regulators	Nuclear receptor coactivator 5; SWI/SNF complex subunit 2.
tRNA modification	Lupus La.
Ribosome biogenesis	60S ribosomal protein L35a; probable ribosome biogenesis protein RPL24.
RNA processing	Exosome complex exonuclease MTR3; ribonuclease P protein subunit p14; U6 snRNA-associated Sm-like protein LSm8.
Complexes I + II + III	
RNA helicases	Helicases A, DDX1, DDX18, DDX24, DDX3X, DDX10, DDX47, DDX49, DDX5, DHX15.
Ribonucleoproteins	HnRNPs – A0, A2/B1, A3, C1/C2, F, H, H2, H3, K, L, M, Q, R, U, U-like protein 2. snRNPs E, Sm D1, Sm D2, Sm D3, U1 RNP A and A', U5 200 kDa helicase, U1 70 kDa, U4/U6 RNP Prp31, 116 kDa U5 component, H/ACA RNP subunit 2 and 4.
Processing factors	Spliceosomal protein SAP 155; SF-3 subunit 1 and 2; SF-3B subunit 3 and 4; U2AF 65 kDa subunit; SF-arg/ser rich 7; SF-13A; CSTF 77 kDa subunit; CPSF subunit 6 and 7.

[a] suggested contaminants.