Interaction Networks of the Molecular Machines That Decode, Replicate, and Maintain the Integrity of the Human Genome*

Benoit Coulombe‡§, Célia Jeronimo‡, Marie-France Langelier‡, Marilena Cojocaru‡, and Dominique Bergeron‡

The interaction of many proteins with genomic DNA is required for the expression, replication, and maintenance of the integrity of mammalian genomes. These proteins participate in processes as diverse as gene transcription and mRNA processing, as well as in DNA replication, recombination, and repair. This intricate system, where the various nuclear machineries interact with one another and bind to either common or distinct DNA regions to create an impressive network of protein-protein and protein-DNA interactions, is made even more complex by the need for a very stringent control in order to ensure normal cell growth and differentiation. A general methodology based on the in vivo pull-down of tagged components of nuclear machines and regulatory proteins was used to study genome-wide protein-protein and protein-DNA interactions in mammalian cells. In particular, this approach has been used in defining the interaction networks (or "interactome") formed by RNA polymerase II, a molecular machine that decodes the human genome. In addition, because this methodology allows for the purification of variant forms of tagged complexes having site-directed mutations in key elements, it can also be used for deciphering the relationship between the structure and the function of the molecular machines, such as RNA polymerase II, that by binding DNA play a central role in the pathway from the genome to the organism. Molecular & Cellular Proteomics 3:851–856, 2004.

PROTEIN-PROTEIN AND PROTEIN-DNA INTERACTION NETWORKS REVEALED THROUGH TAP-TAGGING OF NUCLEAR PROTEINS

The cell nucleus houses the genomic DNA and serves as a factory where many molecular machines, often proteins but also ribonucleo-protein complexes, act on genomic DNA to achieve crucial cell functions. These molecular machines carry catalytic activities necessary to i) transcribe parts of the genome into RNA molecules required for gene expression, ii) replicate chromosomal DNA to generate the genetic material transmitted during cell division, and iii) modify DNA through recombination and repair mechanisms so as to preserve its integrity despite various deleterious conditions. The nucleus also contains many regulatory proteins that modulate the activity of these molecular machineries often by regulating their interaction with specific enzymes or with chromatin DNA. Finally, a large number of nuclear proteins interact with RNA molecules in processes such as splicing, polyadenylation, and capping.

For many years, defining the interactions between the various proteins involved in nuclear function has been a major goal for several reasons. First, interacting proteins are often involved in the same biochemical pathway; defining the interacting partners of a protein of unknown function often helps to infer its function; defining the interacting partners of a known protein is invaluable to detailing its mechanism of action and regulation. Second, the idea that different nuclear processes are sometimes tightly coupled requires that components from different pathways interact at specific times and locations; defining specific interactions between proteins may define mechanisms by which nuclear processes are integrated. Third, the spatial localization of proteins at specific sites within the cell nucleus suggests that key interactions between them and structural proteins are required; defining the association of a nuclear machine or regulatory factor with structural proteins may help in defining novel cell regulatory mechanisms.

The interaction of proteins with specific genomic DNA sequences is also essential to direct the molecular machines and regulatory factors to their sites of action, such as gene promoters for transcription, replication origins, damaged DNA, and so on. The recruitment of molecular machines, such as RNA polymerase II, to specific genomic locations usually involves the action of regulatory proteins, including the transcriptional activators and co-activators that recognize short sequence elements located in promoters upstream of transcription units or in enhancers that can be located at various distances either upstream or downstream of a transcription initiation site.

Over the years, many different techniques have been used to define protein-protein and protein-DNA interactions (1, 2). In many cases, these techniques are performed in the test
tube and use protein and DNA components extracted from their cellular environment. For protein-protein interactions, they include in vitro pull-down and affinity chromatography. In the case of protein-DNA interactions, we can refer to gel mobility shift assays and DNA affinity chromatography. No doubt that the use of these in vitro techniques is necessary to define the molecular details of the interactions as well as the catalytic and regulatory mechanisms resulting from specific interactions. For the screening of novel interactions, however, in vitro techniques are recognized as often generating false positives (3).

Fortunately, techniques for defining both protein-protein and protein-DNA interactions as they prospectively exist in vivo have recently been developed (Fig. 1). In vivo pull-down of i) protein complex assemblies (9) and ii) cross-linked chromatin fragments (10) have been performed in large-scale studies. Yeast was used as a model to show that expressing, at a physiological level, a polypeptide of interest with a double-affinity tag (tandem affinity peptide or TAP)1 (11, 12) is an

---

1 The abbreviations used are: TAP, tandem affinity peptide; ChIP, chromatin immunoprecipitation; RNAPII, RNA polymerase II; GTF, general transcription factors.
efficient tool to purify native protein complexes and to define protein-protein interactions with a reduced number of spurious interactions (Fig. 1, left). By this method, the affinity tag is used to perform tandem affinity chromatography in native conditions and the eluate is analyzed by MS (13). Reciprocal tagging of interacting partners allows validation of the results. Bioinformatics tools have also been developed to integrate the TAP-tagging data with complementary interaction datasets (7, 8, 11, 12). The in vivo pull-down procedure has been shown to reduce to a minimum both the false positives and false negatives when studying protein-protein interactions (7, 8). In the case of protein-DNA interactions, chromatin immuno precipitation (ChIP) experiments using antibodies directed against a nuclear protein can be used to localize this protein on chromatin DNA in vivo (14). In ChIPs, the chromatin is first cross-linked in vivo by treating the cells with formaldehyde and, after cell lysis, the chromatin is fragmented and then used in an immunoprecipitation experiment using a specific antibody. The immunoprecipitated DNA is purified and analyzed using single-locus PCR amplification or, in order to increase the throughput, hybridized to a DNA microarray containing many thousands of DNA sequences, usually gene promoters (15, 16). Of particular interest, TAP-tagged proteins can be used in ChIP experiments in addition to their use in the purification of free protein complexes (Fig. 1, right). In yeast, the so-called TAP-ChIP has been used to determine the location of many proteins along genomic DNA (17, 18). This system can also be used in mammalian cells (19).

AN EXAMPLE: THE HUMAN RNA POLYMERASE II INTERACTOME

RNA polymerase II (RNAPII) is the multi-subunit DNA-dependent RNA polymerase that transcribes protein-coding genes and plays a crucial role in transcriptional control. The RNAPII interactome represents the protein-protein interactions made by RNAPII and the protein-DNA interactions made by some protein components of the interactome (as determined using the TAP procedure described in Fig. 1). The example presented here shows only the major polypeptides found in association with RNAPII and the interaction with two hypothetical genes (A and B). Each distinct color identifies the subunits of a protein (RNAPII, TFIIF, TFIIB, and FCP1), as this protein was previously purified using classical chromatography methods. The TAP is represented by a triangle apposed to the eight polypeptides that were used in this specific analysis. A dashed line with the same color code as the TAP (triangle) groups together the polypeptides found in the corresponding eluate; the color code is also used to link tagged polypeptides to the DNA regions where they associate as determined by TAP-ChIP experiments. For example, the TAP-tagged RPB11 subunit of RNAPII affinity purified a complex containing 17 major subunits. TAP-tagged RPB11 was also found in association with both the promoter and transcribed region of both genes A and B.

Fig. 2. The RNAPII interactome. The human RNAPII interactome represents the protein-protein interactions made by RNAPII and the protein-DNA interactions made by some protein components of the interactome (as determined using the TAP procedure described in Fig. 1). The example presented here shows only the major polypeptides found in association with RNAPII and the interaction with two hypothetical genes (A and B). Each distinct color identifies the subunits of a protein (RNAPII, TFIIF, TFIIB, and FCP1), as this protein was previously purified using classical chromatography methods. The TAP is represented by a triangle apposed to the eight polypeptides that were used in this specific analysis. A dashed line with the same color code as the TAP (triangle) groups together the polypeptides found in the corresponding eluate; the color code is also used to link tagged polypeptides to the DNA regions where they associate as determined by TAP-ChIP experiments. For example, the TAP-tagged RPB11 subunit of RNAPII affinity purified a complex containing 17 major subunits. TAP-tagged RPB11 was also found in association with both the promoter and transcribed region of both genes A and B.
A structure-function analysis of the RNAPII-DNA interaction interface. A, RPB2, the second largest and catalytic subunit of RNAPII, carrying the TAP and mutated in key structural elements within the active center, was expressed in human cells. Wild-type and mutated TAP-RPB2 (a RSR-to-AAA triple amino acid substitution in this example) were used to purify variant forms of RNAPII. B, a model of elongating RNAPII (gray) with DNA (green) showing a sample of the many sites (amino acids in red) that were individually mutated in TAP-tagged RP22. Each of these variants can be produced and tested in various assays. C, the wild-type and mutated enzymes were compared in assays that monitor the various stages of the transcription reaction (RNAPII complex assembly, promoter binding, abortive initiation on premelted promoter, abortive initiation on closed promoter, and elongation). The example summarizes data obtained with RNAPII mutants in Switch 3 and Fork Loop 1 (see text). Signs (+, +/–, –) represent the relative activity of each mutant in the various assays.
genes in eukaryotes. Early studies of mammalian RNAPII transcription have revealed that the enzyme cannot by itself recognize and bind a promoter, melt the DNA around the transcription start site, and initiate RNA chain formation. To do so, it requires a set of accessory factors, the so-called general transcription factors (GTF), including TFIID, TFIIF, TFIIF, TFIIE, TFIIF, and TFIH (20–22). Although these factors were first believed to be required for the transcription of all the protein-coding genes, recent reports indicate that most GTF are required for the transcription of only a set of genes (23, 24). The mammalian GTF have been purified using classical chromatography approaches and their subunits cloned (20–22). The purification of the GTF and RNAPII using classical chromatography methods has been performed such that the proteins are exposed to stringent cell lysis conditions and high salt concentrations. Such conditions have most likely led to the dissociation of some polypeptides, possibly regulatory subunits. Alternatively, numerous reports have described protein-protein interactions between the various GTF, off and on the promoter DNA, as well as between the GTF and other nuclear proteins (20–22). Many of these interactions were defined using in vitro protein affinity chromatography, in vitro pull-down experiments, or yeast two-hybrid assays in conditions that are not physiological.

In the course of a large-scale effort aimed at deciphering both the protein-protein and protein-DNA interactions involved in nuclear events, we have programmed human cells to express near physiological levels of many components of the RNAPII general transcription machinery carrying a TAP tag at their C terminus (19). In each case, protein complexes were purified and their components identified using MS. In many cases, the identity of the polypeptides was confirmed by Western blot and/or functional assays. When the eluted complex contained many polypeptides, we proceeded to tag at least one, but often many additional subunits. This system was very useful for defining transcriptional regulatory networks. Fig. 2 represents a part of the human RNAPII interactome. In this case, eight different polypeptides of the interactome were individually tagged. As defined by our TAP-tagging experiments, the RNAPII interactome contains 17 major protein components that also interact with many gene regions (only two genes, with promoter and transcribed region, are represented in the general example of Fig. 2). The human RNAPII interactome brings together many proteins including: i) all 12 subunits of RNAPII as the enzyme was previously described using classical biochemical procedures (21, 25); ii) the two subunits of TFIIF (RAP74 and RAP30), a factor previously shown to bind to RNAPII and regulate its association with the promoter by stimulating DNA unwrapping around the core of the enzyme (26–28); iii) TFIIB, a factor known to associate with both TFIIF and RNAPII, to recognize promoter DNA, and to interact with the TATA box-binding subunit of TFIID (20, 22); iv) FCP1, a phosphatase that specifically dephosphorylates the CTD domain of RNAPII in order to maintain the enzyme in a form competent for promoter recruitment (29, 30); and v) the RNAPII-associated protein 1 (RPAP1), a novel polypeptide of unknown function that appears to be tightly associated with RNAPII both structurally and functionally (19). The results of our TAP-ChIP experiments reveal that some components of the RNAPII interactome are associated with specific genomic regions (M. Cojocaru and B. Coulombe, unpublished data). In the example shown, which includes two genes expressed in all human cells, RNAPII, TFIIF, and TFIIB have specific patterns of interactions with the promoter and transcribed region of these genes.

A FUNCTIONAL ANALYSIS OF THE INTERFACE BETWEEN RNAPII AND DNA

TAP-tagging of specific mutant polypeptides was also used to isolate variant forms of molecular machines and regulatory proteins (Fig. 3). For example, we have used TAP-tagged Rpb2, the second largest and catalytic subunit of RNAPII, having mutations in specific structural domains found within the active center, to isolate the mutated forms of RNAPII (19). Because the tandem affinity purification procedure is carried out in native conditions, the purified mutants can be analyzed in various functional assays that recapitulate in vitro the various stages of the transcription reaction. ChIP assays have allowed analyzing the ability of the mutated RNAPII to associate with promoter DNA and to transcribe genes in vivo. Guided by the high-resolution structures available for RNAPII (31–33), we have used this procedure to analyze a large number of RNAPII mutants and begin a systematic structure-function analysis of this important enzyme. As an example, Fig. 3 summarizes the results of experiments showing that a motif called “switch 3” is involved in promoter binding by the RNAPII preinitiation complex and a motif called “fork loop 1” appears to modulate the interaction of the polymerase with the transcription bubble and/or short nascent transcripts at the initiation stage of the transcription reaction (19).

The approach described here can be applied to all the proteins interacting with the genome. Its use will produce a wealth of novel information on both the regulation and mechanisms of action of the molecular machines that decode, replicate, and maintain the integrity of the genome.

Acknowledgments—We are grateful to the members of our laboratory and Jacques Archambault and François Robert for helpful discussions. We thank Julie Edwards for critical reading of the manuscript and Diane Bourque for artwork.

* This work was supported by grants from the Canadian Institutes for Health Research, Genome Canada, and Genome Québec. B. C. is a senior scholar from the Fonds de la Recherche en Santé du Québec. C. J. holds a studentship from the Canadian Institutes for Health Research and M.-F. L. from the Natural Sciences and Engineering Research Council of Canada and the Fonds Québécois de la Recherche sur la Nature et les Technologies. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.
REFERENCES


