Structural heterogeneity in the intrinsically disordered RNA polymerase II C-terminal domain

Bede Portz1, Feiyue Lu1,2, Eric B. Gibbs3, Joshua E. Mayfield4, M. Rachel Mehaffey5, Yan Jessie Zhang4,6, Jennifer S. Brodbelt5, Scott A. Showalter1,3 & David S. Gilmour1

RNA polymerase II contains a repetitive, intrinsically disordered, C-terminal domain (CTD) composed of heptads of the consensus sequence YSPTSPS. The CTD is heavily phosphorylated and serves as a scaffold, interacting with factors involved in transcription initiation, elongation and termination, RNA processing and chromatin modification. Despite being a nexus of eukaryotic gene regulation, the structure of the CTD and the structural implications of phosphorylation are poorly understood. Here we present a biophysical and biochemical interrogation of the structure of the full length CTD of Drosophila melanogaster, which we conclude is a compact random coil. Surprisingly, we find that the repetitive CTD is structurally heterogeneous. Phosphorylation causes increases in radius, protein accessibility and stiffness, without disrupting local structural heterogeneity. Additionally, we show the human CTD is also structurally heterogeneous and able to substitute for the D. melanogaster CTD in supporting fly development to adulthood. This finding implicates conserved structural organization, not a precise array of heptad motifs, as important to CTD function.
NA polymerase II (Pol II) transcribes primarily protein-coding genes in eukaryotes, a process regulated at initiation, elongation and termination, and concomitant with chromatin modification and RNA processing. The temporospatial regulation of transcription and co-transcriptional processes involves the concerted action of multiple factors bound to Pol II via a long, repetitive, intrinsically disordered C-terminal domain (CTD) comprised of heptads of the consensus sequence YSPTSPS. In a process dubbed the CTD code, each non-proline residue of the repeat can undergo cis–trans isomerization, creating a multitude of distinct binding sites for CTD-interacting proteins.

Despite its role as a regulatory nexus, little is known about the structure of the CTD, and most studies have focused on the structure of short CTD peptides. Such peptides have been shown to adopt β-turn structures when bound to CTD-interacting proteins, and β-spiral models of the CTD have been proposed that extrapolate β-turns across the entire length of the CTD. However, nuclear magnetic resonance (NMR) and circular dichroism studies of CTD peptides suggest such turns are rare among the ensemble of CTD structures, and only populated to a high degree in turn-promoting solvents at low pH or in the context of circularized peptides. Phosphorylation enables formation of additional hydrogen bonds that have been shown to contribute to turn structures in CTD peptides bound to factors, but phosphorylation fails to increase the β-turn propensity of unbound CTD peptides in solution. Together, these data argue against models involving turns across the full CTD in either the apo or phospho state. Such models are entropically unlikely, as they assume the CTD simultaneously adopts multiple low-probability turns. Additionally, β-spiral models fail to account for local structural variation potentially imparted by those heptads deviating from the consensus sequence, and such repeats comprise the majority of the CTD in developmentally complex organisms.

Existing structural information describing the full-length CTD is limited. The CTD, likely due to its flexibility, is absent from X-ray crystal structures of Pol II, but one study showed it may share space with the CTDs of adjacent Pol II molecules in the crystal, a space too small to accommodate extended structures, leading to a model that was sufficiently compact to fit into the space provided in the crystal, yet lacking the order associated with folded, globular structures. The CTD is also absent from cryoEM derived models of elongating Pol II, again likely due to its flexibility. Additional evidence for a compact CTD was obtained from cryoEM models of a GST-CTD fusion protein bound to the middle module of the multi-subunit transcriptional regulator known as the Mediator complex. However, it is unknown to what extent binding to Mediator and possible CTD:CTD interactions resulting from GST-CTD dimers contribute to the observed compaction. Earlier studies of negatively stained two-dimensional (2D) crystals of Pol II and Pol II mutants lacking the CTD revealed density differences attributed to a flexible CTD and linker occupying a space measuring ~100 Å in diameter. The extent to which fixation on a solid surface alters the CTD ensemble in the 2D crystal studies is unknown, and like other X-ray crystallography studies, the proximity of CTDs from adjacent Pol II molecules in the 2D crystals could potentially alter the CTD ensemble through CTD:CTD interactions. One study by Corden and Zhang examined the structure of the mammalian CTD in solution and found the hydrodynamic radius of the CTD extends as a function of phosphorylation. Consistent with extension was the observation by Laybourn and Dahmus that the CTD is more proteolytically labile in the phosphorylated state, suggesting conformations more accessible to CTD:protein interaction. However, it is unclear from this work if the CTD was rendered more accessible by CTD phosphorylation alone, or in concert with possible interactions with proteins present in the HeLa extract component of the phosphorylated Pol II cleavage experiments, but absent from the unphosphorylated Pol II reactions. Nevertheless, evidence supports an unphosphorylated CTD existing as a compact structural ensemble, capable of extending as a function of phosphorylation. Such an ensemble is consistent with emerging structural understanding of intrinsically disordered proteins (IDPs), which can be more compact than chemically denatured proteins of the same size, and can be structurally heterogeneous, possessing transient structural features.

The notion that IDPs can adopt semi-compact, heterogeneous ensembles is important in the context of CTD evolution. Much of our knowledge comes from studies on the CTD of Saccharomyces cerevisiae that is unique in its sequence homogeneity, with nearly every heptad exactly matching the YSPTSPS consensus. The CTD in mammals contains a consensus region proximal to the catalytic core of Pol II, and a non-consensus region at the distal end of the CTD composed of heptads differing primarily at the seventh position. (Fig. 1a,b). The CTD of Drosophila melanogaster is distinct from its human counterpart in that only two of forty-two heptads exactly match the consensus, though a similar consensus motif is identified. Evolutionary differences in CTD composition raise the possibility that different motifs may serve to organize CTD:protein interactions via the co-evolution of factors that favourably recognize particular CTD motifs. An alternative and not mutually exclusive role for non-consensus motifs could be the formation of structural heterogeneity, rendering particular regions of the CTD more accessible to protein interaction.

A model of CTD structure derived from data collected on the monomeric, full-length, metazoan CTD, under physiologically relevant buffer conditions, that relates global and local structure has been lacking. Here we present a biophysical and biochemical characterization of the CTD of D. melanogaster, which we identify as a structural homologue to its human counterpart. Using a combination of size exclusion chromatography (SEC) and Small Angle X-ray Scattering (SAXS), we show that the CTD undergoes moderate extension as a function of phosphorylation by the CTD kinase P-TEFb and a concomitant decrease in flexibility. In spite of the repetitive nature of the CTD sequence, we find structural heterogeneity in the CTD that is maintained in the phosphorylated state. Despite sequence differences between the CTD of D. melanogaster and H. sapiens, the human CTD is structurally similar to the fly CTD globally and both are structurally heterogeneous locally. Surprisingly, the human CTD is capable of functioning in place of the fly CTD in vivo, supporting the development of flies to adulthood. This finding implicates conserved structural organization, not a precise series of heptad motifs, as important to CTD function.

**Results**

A similar global CTD structure exists despite sequence differences. CTD length and sequence complexity differ among organisms, with non-consensus heptads predominating in the longer CTDs of metazoans. Nevertheless, aligning blocks of seven amino acids beginning with YSP for the CTDs of D. melanogaster and H. sapiens using Weblogo returns a similar consensus motif with most deviation occurring in the seventh position. (Fig. 1b). We sought to determine whether the presence of non-consensus repeats impacts the global structure of the CTD. Because the CTD naturally exists as a terminal extension emanating from the globular and acidic Pol II, and to
aid in purification and solubility, we expressed and purified the CTDs of *S. cerevisiae*, *D. melanogaster*, and *Homo sapiens* as maltose binding protein (MBP) fusions and compared their mobility using native-PAGE. The relationship between CTD length and mobility for the three species suggests gross structural similarity despite differences in the relative proportion of non-consensus heptads (Fig. 1c).

To quantify the relationship between CTD length and overall structure more precisely, we compared the three CTDs using size exclusion chromatography (SEC), determining the Stokes radii ($R_s$) using a standard curve derived from natively folded proteins, and found a linear relationship between $R_s$ and molecular weight (MW) for the CTD fusions. At some minimum CTD length, the MBP portion of the fusion could dominate the analysis, and it could be expected that this linear relationship would break down, confounding interpretation of the results. Thus, we included in this analysis a truncated version of the fly CTD, dubbed W81, shown to have functional defects in vivo. Despite removing 50% of the CTD, generating a CTD shorter than that of yeast, this fusion also remained in the linear range of the CTD curve (Fig. 1d). The relationship between $R_s$ and MW for all of the CTD fusions suggests that they are structural homologues despite sequence differences, sharing an extended conformation compared to...
natively folded proteins of similar MW (Supplementary Fig. 1a,b). The correlation between \( R_g \) and MW for the CTDs creates a CTD standard curve, with the *Drosophila* CTD at the center, that can be used to interrogate structural changes resulting from phosphorylation.

**Phosphorylation by P-TEFb extends the CTD.** Phosphorylation of the *D.mel*CTD by the CTD kinase P-TEFb lead to an apparent ~26% increase in the Stokes radius of the CTD, causing the phospho-CTD to fall off the CTD standard curve (Fig. 1d). To further interrogate differences in the structures of the apo and phospho CTD, we employed small angle X-ray scattering (SAXS), a high-precision technique well suited to the interrogation of IDPs in solution\(^{32} \). The CTD was subject to extensive phosphorylation and subsequent gel filtration chromatography, resulting in homogenous preparations of apo and phospho species with discrete mobility in SDS-PAGE and with no evidence of unphosphorylated protein remaining in the phospho-CTD sample (Fig. 2a). Using MALDI-TOF mass spectrometry, we determined that phosphorylation by P-TEFb under our conditions lead to an increase in MW of our MBP-*D.mel*CTD fusion of ~1.9 kDa, equivalent to the mass of ~25 phosphates (Supplementary Fig. 2a,b). Tandem mass spectrometry localizes this phosphate incorporation to primarily the fifth position, unless a heptad lacks a SP motif at positions 5–6 (Supplementary Fig. 2c), consistent with the *in vitro* preference for serine 5 for P-TEFb observed previously\(^{26} \). The density of phosphorylation; less than one phosphate per heptad with no heptads phosphorylated more than once, is in good agreement with *in vivo* observations, suggesting our biophysical analyses focus on a biologically relevant CTD phospho-isofrom\(^{27,28} \). Guinier analysis of SAXS curves obtained across a range of concentrations for the apo and phospho CTD in good agreement, showing no concentration effects across the dilution series in measured MW or radius of gyration (\( R_g \)), with phosphorylation leading to a ~15% increase in \( R_g \) (Supplementary Fig. 3)\(^{29,30} \) and an increase in MW of ~1.9 kDa, in excellent agreement with the MALDI-TOF data (Supplementary Fig. 2a,b). Pairwise distance distribution functions for the averaged scattering curves from the apo and phospho MBP-*D.mel*CTD fusions were generated using AUTOGNOM (Fig. 2b)\(^{31} \). The \( R_g \) values obtained via the \( P_r \) versus \( r \) analysis are slightly larger than those obtained via Guinier analysis, which is not uncommon for IDPs, and the overall extension calculated in this analysis is in good agreement with that obtained from the Guinier fit (Fig. 2b)\(^{32} \). These analyses provide independent validation of the phosphorylation dependent extension we observe using SEC (Fig. 1d).

**The CTD is semi-compact.** The Flory chain model applied to disordered polypeptides relates protein length to \( R_g \) via a power law, and has been used as a benchmark against which to compare the disorder, or unfoldedness, of a protein\(^{33} \). Applied only to the CTD portion of the fusion protein and omitting MBP, the Flory model for denatured proteins predicts an \( R_g \) of ~60 Å (ref. 30). Notably, this is larger than the measured \( R_g \) of 51.8 Å for the entire MBP-*D.mel*CTD fusion, which includes MBP, a globular protein with a mass exceeding that of the CTD and an estimated \( R_g \) of ~22 Å. This comparison suggests that the CTD is considerably more compact than a denatured coil. Together, our measurements support a semi-compact CTD, more extended than a globular protein of the same molecular weight, but more compact than a denatured coil.

**Ensemble optimization models of the CTD.** To better conceptualize the CTD structure on the global scale, we generated ensemble models from the SAXS curves collected on apo and phospho MBP-*D.mel*CTDs using the Ensemble Optimization Method (EOM) 2.1, accounting for the contribution of the globular MBP\(^{24,34} \). This method generates 10,000 models for the CTD that are parsed using an algorithm to yield an ensemble whose average scattering curve agrees with the experimentally measured curve for our apo and phospho MBP-*D.mel*CTD fusions. These models yielded ensembles for the apo and phospho MBP-*D.mel*CTD with \( R_g \) values of 50.91 and 60.98 Å, which are in good agreement with \( R_g \) values derived from the \( P(r) \) versus \( r \) analysis (Figs 2b and 3). The resulting ensembles are depicted oriented by MBP and displayed with the 12-subunit RNA Pol II elongation complex model to allow for size comparison (Fig. 3)\(^{35,36} \). Notably, both compact and extended structures are found in the ensembles describing both the apo and phospho CTD, with a bias towards extended structures in the phospho state consistent with the extension observed in SEC and SAXS. SAXS data lack the resolution to interpret the models on the amino-acid or heptad scale, yet do provide a visual representation of the scale of the monomeric CTD in solution that is constrained by and in agreement with experimental data.

**The fly and human CTDs are structurally heterogeneous.** That the CTD is semi-compact and capable of undergoing extension as a function of phosphorylation prompted us to examine the structure at a more local level. Limited proteolysis experiments have been used to interrogate IDP structure, with regions of protease sensitivity and protection reflecting more unstructured and structured regions, respectively\(^{37–39} \). The repetitive nature of the CTD is ideally suited to this approach, as predicted cleavage sites are evenly distributed across the length of the CTD. If the CTD is either a completely disordered chain, or if it is structurally repetitive as β-spiral models predict, proteases would be predicted to cleave the CTD with equal probability along its length. Alternatively, local structural heterogeneity would manifest as regions of protease sensitivity or protection that can be mapped to regions of the CTD by comparison to a CTD ladder (Fig. 4a and Supplementary Fig. 4a–c). Chymotrypsin cleaves the peptide bond C-terminal to tyrosine and is predicted to cut the *Drosophila* CTD 43 times across its length\(^{40} \). Surprisingly, chymotrypsin yields a pattern of proteolysis characterized by a hypersensitive site towards the distal part of the CTD, and sensitivity located across the proximal region, with a protease hypersensitive site towards the distal part of the CTD, and sensitivity located across the proximal region, with a protease hypersensitive site towards the distal part of the CTD, and sensitivity located across the proximal region.
insensitive region spanning the central part of the CTD that is cut in only one location (Fig. 4b). One interpretation of this result is that the CTD is not uniformly accessible to protein-sized factors due to local structural variation. Alternatively, this result could be due to sequence variation in the Drosophila CTD giving rise to non-uniform recognition of potential cleavage sites. To distinguish between these hypotheses, we subjected the MBP-D.mel.CTD to limited proteolysis with subtilisin, a protease possessing similar specificity to chymotrypsin but that is structurally heterogeneous across its length. (a) Experimental design: end-labelled MBP-D.mel.CTD is subject to limited proteolysis, SDS-PAGE, and autoradiography. A completely unstructured, or alternatively, a structurally repetitive CTD is predicted to generate a uniform pattern of proteolytic fragments. A structurally heterogeneous CTD is predicted to give rise to a non-uniform pattern. (b) Limited proteolysis with chymotrypsin generates a non-uniform pattern of CTD fragments, with a hypersensitive site in the distal CTD, a sensitive region in the proximal CTD, and a largely protease insensitive region in the central CTD that is cleaved at only one site. The right most lane contains radiolabelled CTD fusion proteins with the molecular weights designated on the right of the panel. (c,d) Limited proteolysis with subtilisin or proteinase K reveal similar sites of sensitivity and protection. (e) A direct comparison of proteolytic fragments generated by three proteases shows a similar by not identical pattern of proteolysis. (f) Sodium dodecyl sulfate (SDS) alters the relative proteolytic sensitivity of the CTD to proteinase K, enhancing the sensitivity of the CTD at sites that generate bands near 35 kD and below 19 kD relative to the CTD ladder (0.1% SDS lane). 0.5% SDS renders the globular MBP portion of the fusion protein largely protease insensitive region spanning the central part of the CTD that is cleaved at only one site. The right most lane contains radiolabelled CTD fusion proteins with the molecular weights designated on the right of the panel. (g) The human CTD is radiolabelled on the final acidic repeat by casein kinase II (CK2) (uncut lane). TEV cleavage to separate the human CTD from the MBP fusion demarcates the point below which protease sensitivity occurs in the CTD portion of the protein (TEV lane). Limited proteolysis with proteinase K reveals a distal hypersensitive site reminiscent of that observed in the Drosophila CTD (compare to D.mel.CTD ProK lane), a protease hypersensitive proximal site near MBP and a central region that is largely protease insensitive. Both the human and fly CTDs share distal protease hypersensitivity, proximal sensitivity that is localized to a discrete region in the human CTD, and a more central region of the CTD that is less protease sensitive.
and human, with the human CTD also exhibiting a hypersensitive site in the distal region, in addition to a hypersensitive site in the proximal region near where the collection of sensitive sites resides in the Drosophila CTD, suggesting that conservation in CTD structure between fly and human exists despite length and sequence differences (Fig. 4g).

**Structural heterogeneity persists after phosphorylation.** In light of the findings that the CTD undergoes extension as a function of phosphorylation and is structurally heterogeneous in the apo state, we sought to determine if the CTD is locally reorganized in the phospho state. To address this question, we performed limited proteolysis on the phospho CTD using proteinase K (Fig. 5a). The phosphorylated CTD retains a non-uniform pattern of cleavage with the resulting proteolytic fragments having shifted mobility due to phosphorylation, indicating local structural heterogeneity is maintained (Fig. 5a). However, the sensitivity of the various regions changes relative to each other. For example, at digestion points with minimal amounts of protease, products in the lower region of the gel are more abundant than products near the central region for the apo CTD (Fig. 4d) whereas they are nearly equivalent in the case of the phosphorylated CTD (Fig. 5a). Overall, phosphorylation of the CTD results in more even level of cutting at all of the labile sites.

**CTD phosphorylation increases protein accessibility.** The more even cleavage of protease accessible sites across the phospho CTD prompted us to test whether phosphorylation increases CTD accessibility. We subjected a mixture of apo and phospho CTD to limited proteolysis and compared the level to which each isoform was digested as a function of increasing protease concentration (Fig. 5b). We quantified three experimental replicates, comparing the amount of intact CTD remaining at each protease concentration. This revealed a moderate increase in the level of proteolysis of the phospho CTD, which is digested past the point of single hit kinetics (>50% cleavage) at a lower protease concentration, consistent with a moderate increase in accessibility as a function of phosphorylation (Fig. 5c).

**The CTD stiffens as a function of phosphorylation.** SAXS data can be used to make qualitative comparisons of the flexibility of proteins and to monitor flexibility changes after post-translational modification. Kratky plots enable comparison of protein flexibility, and indicate that the CTD becomes less flexible when phosphorylated (Fig. 6a and Supplementary Fig. 5). To further validate this result we generated Porod–Debye and Kratky–Debye plots (Fig. 6b,c) from the apo and phospho CTD scattering curves. These analyses truncate the high q region of the data and thus are less prone to potential artifacts of buffer subtraction. The asymptotic rise in the Porod–Debye plot exhibited by the apo CTD curve is characteristic of flexible proteins and is diminished for the phospho CTD (Fig. 6b). Further, the plateau observed for the apo CTD in the Kratky–Debye plot is lost for the phospho CTD (Fig. 6c). The divergent behaviour of the apo and phospho CTD between the Kratky–Debye and Porod–Debye plots and the differences within each plot for the apo and phospho CTD are indicative of a phosphorylation-dependent decrease in flexibility.

**The human CTD can function in place of the fly CTD in vivo.** The structural similarity but sequence differences between the fly and human CTDs, coupled with the genetic tractability and developmental complexity of Drosophila, emboldened us to test the hypothesis that the evolution of non-consensus CTD repeats imparted important lineage specific functionality. The high degree of conservation of the CTD sequence among the twelve species of Drosophila (Fig. 7a) and separately among mammals contrasts with the sequence divergence between the two phyla (Fig. 1a). This suggests that within various lineages, specific CTD sequences arose and were maintained under selective pressure, perhaps co-evolving with CTD interacting factors in a way that could recruit new activities to RNA Pol II or organize multiple CTD:factor interactions across the CTD. To test the importance of a precise CTD amino acid sequence in a developmental context, we used RNAi to knock down endogenous Rpb1 in the developing fly wing and determined whether coincident ectopic expression of an RNAi resistant Rpb1 with either the normal Drosophila CTD or the human CTD

Figure 5 | Local structural heterogeneity is maintained in the phospho CTD but the phospho CTD is more accessible to protein interaction. (a) Limited proteolysis of the phospho CTD shows an altered pattern of proteolysis. The hypersensitivity of the distal site is reduced, and the pattern of proteolysis is more evenly distributed among the sensitive sites. The proximal sensitive region, and central protected region are preserved after phosphorylation, with proteolytic fragments shifted relative to the apo fragments due to phosphorylation across the CTD. (b) Representative limited proteolysis of a mixture of apo and phospho CTD to compare protease accessibility. (c) Average of three replicates of the apo (blue bars) and phospho (red bars) mixing experiments quantified as the percentage of intact CTD remaining at each protease concentration. 100–50% intact CTD is the single hit kinetics range of the experiment. Error bars depict s.e.m.
would revert phenotypes caused by depleting the endogenous Rpb1. RNAi-mediated depletion of Rpb1 in the wing caused a wing undergrowth phenotype that could be reverted to wild-type by co-expressing RNAi resistant *Drosophila* Rpb1 (Supplementary Fig. 6a,b). Coincident expression of a defective version of *Drosophila* Rpb1 called W81, which has approximately half of the CTD deleted from the region distal to the body of Pol II, only partially restored wing growth, showing that the assay is sensitive enough to reveal a gradient of phenotypes (Supplementary Fig. 6c). Remarkably, expression of RNAi resistant Rpb1 harbouring the human CTD fully reverts the undergrowth phenotype caused by the Rpb1 RNAi (Supplementary Fig. 6d).

Motivated by the results in the wing, we sought to determine whether RNAi-resistant versions of wild-type or humanized Rpb1 could rescue the lethality caused by ubiquitously expressing Rpb1 RNAi in the entire animal and throughout development. In these experiments, flies ubiquitously expressing the transcriptional activator GAL4 under the control of the actin gene promoter are mated to flies harbouring both RNAi against Rpb1 and an additional Rpb1 derivative rendered resistant to RNAi via synonymous mutations, both of which are expressed in the presence of GAL4 (Fig. 7b). Progeny not expressing GAL4, and thus not expressing Rpb1 RNAi, have curly wings (Fig. 7b). Progeny expressing GAL4, and thus expressing Rpb1 RNAi and an RNAi resistant derivative, have straight wings. As expected, approximately 50% of the progeny had straight wings when the Actin-GAL4 driver was crossed to *yw* control flies lacking Rpb1 RNAI (Fig. 7c, *yw*). In contrast, none of the progeny had straight wings when the Actin-GAL4 driver was combined with Rpb1 RNAI alone, indicating lethal levels of Rpb1 knockdown are achieved using this system (Fig. 7c, Rpb1i). In addition, no straight-winged progeny were present when an RNAi-sensitive version of *Drosophila* Rpb1 was over-expressed along with the Rpb1 RNAI (Fig. 7c, Rpb1wt [sens],Rpb1i) indicating that the level of RNAi knockdown is robust enough to kill the flies even when Rpb1 is ectopically overexpressed in addition to endogenous levels. In stark contrast, approximately 50% of the progeny have straight wings when an RNAi-resistant form of *Drosophila* Rpb1 is present, indicating rescue of Rpb1 RNAI by the ectopically expressed derivative (Fig. 7c, Rpb1wt [res], Rpb1i). Remarkably, the RNAi-resistant Rpb1 harbouring the human CTD also rescues the lethality caused by the Rpb1 RNAI (Fig. 7c, Rpb1hu, Rpb1i). Molecular analysis of the straight winged progeny shows that the RNAi reduces the endogenous Rpb1 mRNA by about 20 fold relative to control flies (Supplementary Fig. 7a). Moreover, comparable levels of each ectopically expressed FLAG-tagged Rpb1 are being expressed (Supplementary Fig. 7b,c).

We next sought an independent test of the ability of the human CTD to function in place of the fly CTD that does not rely on RNAi knockdown. To this end, we employed the Rpb1<sup>G0040</sup> fly line, which contains a P-element insertion in the *Rpb1* transgene so that there are crosses are expressing FLAG-tagged Rpb1 derivatives and RNAi, immunostaining chromosomes obtained from multiple larvae generates instances where chromosomes staining positively for FLAG-Rpb1 were located next to chromosomes from individuals not expressing the FLAG-Rpb1 and Rpb1 RNAi. The latter fail to stain with FLAG antibody but are visible using DNA stain (Fig. 7d, left and middle panels). The positive FLAG staining shows that the ectopically expressed Rpb1 derivatives harbouring either the wild-type or the human CTD associate with chromosomes (Fig. 7d, middle panels). Higher magnification images of chromosomes staining positively for FLAG-Rpb1 show that the ectopically expressed wild-type and humanized Rpb1 co-localize on chromosomes with the Rpb3 subunit of Pol II (Fig. 7d, right panels). Together, these data indicate that ectopically expressed Rpb1 with the human CTD is expressed, co-localizes on chromosomes with another subunit of Pol II, and is able to support growth and development when endogenous Rpb1 is knocked down by ~95% to lethal levels.

We next used immunostaining of polytene chromosomes from third instar larvae to further validate that the humanized Rpb1 is functioning normally. Because only ~50% of offspring from our crosses are expressing FLAG-tagged Rpb1 derivatives and RNAi, we recombined our *Rpb1* transgenes with a *daG32-Gal4* transgene so that there would be ubiquitous expression of the *Rpb1* transgene. In triplicate, eight virgin females from the lethal Rpb1 mutant line were crossed to eight males ectopically expressing Rpb1, or to *yw* as a negative control (Fig. 7e and Supplementary Table 1). Each cross yields three possible types of males, which can be distinguished phenotypically (annotated A, B and C in Fig. 7e). ‘A’ males carry a wild-type endogenous Rpb1 allele (Rpb1<sup>+</sup>) on FM7c and are identified by the narrow eye phenotype (B<sup>+</sup>). ‘B’ males result from ectopic expression of humanized (Rpb1<sup>hu</sup>) or wild-type (Rpb1<sup>wt</sup>) Rpb1 complementing the early embryonic lethality of the Rpb1<sup>G0040</sup> allele, and are identified by normal pigmented bodies (y<sup>+</sup>). The presence of B males from both the human and wild-type CTD crosses establishes that these derivatives of Rpb1 can complement the early embryonic lethal allele (Fig. 7e, B offspring). ‘C’ males are the result of non-disjunction and are characterized by yellow bodies (y<sup>-</sup>). The crosses to *yw* failed to complement (Fig. 7e, B offspring) and
Figure 7 | The human CTD can function in place of the Drosophila CTD in vivo. (a) CTD amino acid sequence alignment from 12 species of Drosophila. (b) Experimental design of the human CTD rescue experiment. UAS-Rpb1i is a GAL4 inducible transgene that produces an RNAi against endogenous Rpb1. UAS-Rpb1wt is a GAL4 inducible transgene that encodes a FLAG-tagged, RNAi-resistant derivative of Rpb1. Fly images were created with Genotype Builder5. (c) Results of crosses to test if ectopically expressed derivatives of Rpb1 rescue the lethality caused by ubiquitously expressing Rpb1 RNAi. Each bar on the histogram indicates the percentage of progeny with straight wings. yw: ActGAL4/CyO × yw (n = 137 progeny); Rpb1i: ActGAL4/CyO × UAS-Rpb1i (n = 86 progeny); Rpb1wt [res], Rpb1i: ActGAL4/CyO × UAS-Rpb1 wt, UAS-Rpb1[res] (n = 102 progeny); Rpb1wt [res], Rpb1i: ActGAL4/CyO × UAS-Rpb1 wt, UAS-Rpb1[res], Rpb1wt (n = 91 progeny); Rpb1tu, Rpb1i: ActGAL4/CyO × UAS-Rpb1 tu, UAS-Rpb1 tu (n = 78 progeny). (d) Immunofluorescence of polytene chromosomes from salivary glands of third instar larvae. Larvae were derived from mating ActGAL4/CyO and UAS-Rpb1 wt, UAS-Rpb1 parents, so half of the larvae ectopically expressed Drosophila or humanized Rpb1 (Left and Middle panels). FLAG Rpb1 detected on chromosomes co-localizes with the Rpb3 subunit of Pol II (Right panel). Arrows highlight chromosomes from individuals not expressing FLAG Rpb1 variants (CyO+/−; UAS-Rpb1 wt/−). (e) Genetic complementation assay mating scheme and results. Complementation was scored in males, which have a single copy of the X chromosome on which the endogenous Rpb1 gene resides. Female flies carrying the early embryonic lethal Rpb1 mutant allele, G0040 are crossed to male flies expressing Rpb1wt or Rpb1tu under the control of da-GAL4, or to yw control flies. Crosses were carried out in triplicate with 8 males and 8 female parents per cross and flies were grown at 24°C and 70% humidity. Shown in e are the total counts of male progeny; counts from individual vials are shown in Supplementary Table 1. ‘B’ males indicate that the ectopically expressed Rpb1 complements of the lethal allele. We observe no ‘B’ flies from the yw control crosses.
never yielded ‘B’ offspring, as it lacked an ectopically expressed Rpb1 transgene to complement the embryonic lethal Rpb1G0040 allele, showing that the G0040 allele confers lethality in this assay. These results show that the human CTD can function in place of the highly conserved Drosophila CTD when the endogenous polymerase is present at levels too low to support viability. This finding has two important implications. First, the results argue against the requirement for a precise CTD sequence even in a complex developmental context. Second, the results argue that the conserved structural organization of the fly and human CTD is important for CTD function.

Discussion

Existing models of CTD structure were generated by iterating structural motifs across the length of the CTD. These models leveraged the best available structural data from CTD peptides, but evidence for turn motifs in unbound CTD peptides in physiologically relevant buffer conditions was lacking. Evidence for a compact, but flexible CTD was more compelling, but was potentially biased by CTD:CTD or CTD:factor interactions or fixation conditions required by the various experiments. Here we present biophysical data on the monomeric full-length CTD obtained with both size exclusion chromatography and small angle X-ray scattering that supports a model for CTD structure that we describe as a compact random coil: more extended than predicted for a natively folded protein of the same mass, but significantly more compact than a denatured coil.

The primary role of the CTD is thought to be that of a scaffold protein capable of recruiting proteins to Pol II involved in transcription initiation, elongation, termination, and co-transcriptional processes such as mRNA capping, splicing and polyadenylation; further, these interactions occur in a temporospatially regulated manner. Compactness may protect the CTD from imprecise post-translational modification and factor recruitment out of phase with the transcription cycle. Moreover, compactness could also facilitate the transcription process by preventing the CTD from becoming entangled with the DNA, RNA or other proteins. Using a series of proteases as protein scale probes, we interrogated the structure of the CTD and found a pattern of sensitivity and protection we interpret as evidence of a heterogeneous CTD structure containing regions with varying degrees of disorder. We favour this interpretation over the alternative hypothesis that variation in the primary amino acid sequence of the CTD gives rise to discontinuous proteolysis. First, three separate proteases with variations in structure and specificity all recapitulate a similar but not identical pattern of proteolysis, and these shared patterns result from proteolysis at only a subset of the total cleavage sites predicted by the specificity of each enzyme (Fig. 4b–e). This suggests that each protease first interacts with the accessible regions of the CTD, before cleaving at subtly different sites therein as a function of their individual specificities. Second, the denaturant SDS, known to stimulate proteolysis through the disruption of protein structure, alters the relative pattern of proteolysis of the CTD by proteasome K (Fig. 4f). That some elements of the proteolytic sensitivity are altered by a denaturant suggests we are detecting a co-evolution of non-consensus heptads and CTD binding proteins that impart linear sequence functional diversity, via added specificity of CTD binding interactions originating from the co-evolution of the CTD with its binding partners.

Such specificity could in theory serve to organize multiple proteins on a single CTD or modulate binding specificities among competing factors. Our ability to replace the D. melanogaster CTD sequence, which is highly conserved among Drosophila species, with the human CTD sequence bearing different heptads but a similar structure, suggests the CTD sequence is not organizing binding interactions in an essential way, even in a developmentally complex context (Fig. 7c–e). Instead, we posit that structural organization may serve to preferentially expose particular regions of the CTD that in turn could control the placement of post-translational modifications and organize binding interactions.

Phosphorylation leads to extension of the CTD that is modest as a percentage of its initial size (Figs 1d and 2), which results in modestly increased protein accessibility as measured by a comparison of proteolytic sensitivity between the apo and phospho protein (Figs 4 and 5). Analysis of the SAXS data indicates that the phosphorylated CTD is less flexible than the unphosphorylated CTD (Fig. 6). Meanwhile, local structural heterogeneity is maintained (Fig. 5a). These observations are interesting in the context of those CTD:CTD co-crystal structures of factors that favour binding phosphorylated CTD peptides without directly recognizing the phosphate via hydrogen bonding. Together these observations support a model for CTD binding in which multiple effects of phosphorylation may in concert facilitate factor binding. First, the CTD transitions from a polyanpholoyte to a polyanion, altering its affinity for factors, potentially as a function of phosphate density. Extension de-protects the CTD, making it more accessible to binding. A less flexible CTD with reduced conformational entropy may provide a more binding-competent dock by depleting the CTD ensemble of conformers incompatible with binding, thereby reducing the entropic penalty for binding. A compelling paradigm for phosphoCTD recognition may be the phospho-protein Sic1, which like the CTD is compact, intrinsically disordered, comprised of a repetitive and low complexity sequence, and is subject to phosphorylation at multiple repeats giving rise to multivalent binding. Cdc4 can recognize phospho Sic1 on any one of multiple phosphorylated repeats, but only after a certain threshold of phosphorylation is reached that achieves adequate charge density and possibly reduced conformational entropy which results in binding ‘ultrasensitivity’, binding characterized by extremely high Hill coefficients. Likewise, analogous changes in the CTD could tip the balance in favour of binding and explain how a CTD dynamically sampling extended structures avoids unregulated modification and premature binding in the presence of hundreds of interacting proteins in the nuclear compartment.

Methods

Sequence motif generation. The amino acid sequences of the Pol II CTDs of D. melanogaster and H. sapiens were entered into WebLogo as chains of seven amino acids, beginning with YSP. For compatibility with the tool, repeats longer than seven amino acids were trimmed, those shorter were extended with ‘X’.

Cloning MBP-CTD plasmids. PCR amplicons encoding the S. cerevisiae, D. melanogaster, and H. sapiens CTD with either 6His or FLAG-STREP (DYKDDDDK-WSHPQFEK) tags were inserted into the Xbal digested pMalX vector using T4 ligase to yield the plasmids

Expression of MBP-CTD proteins. pMalX vectors containing the CTD fragments with either C-terminal 6His tags or C-terminal FLAG-STREP (DYKDDDDK-WSHPQFEK) tags were transformed into BL21DE3 E. coli cells and grown in LB
media with 0.2% dextrose and 100 μg ml⁻¹ ampicillin at 37 °C shaking at 250 r.p.m. to an optical density at 600 nm of ~0.5. Protein expression was induced with 300 μM IPTG, and proteins were expressed overnight at 31 °C with shaking at 250 r.p.m. Cells were collected by centrifugation for 10 min at 3,000,000 g, flash frozen and stored at −80 °C. Thawed pellets were lysed by sonication in 20 ml per 11 culture volume of Binding Buffer (50 mM HEPES pH 7.5, 300 mM NaCl, 10% glycerol, 1% NP-40, 2.5 mM imidazole, 2.5 mM β-mercaptoethanol) plus protease inhibitors.

Purification of MBP-CTD proteins. Cells lysates were centrifuged at 100,000 x g for 30 min and the clarified lysates were bound to TALON superflow resin (Clontech) for 6His tagged proteins or Streptactin superflow resin for FLAG-STREP (Iba) in batch for 30 min at 4 °C. Resin was centrifuged at 500 g for 5 min and washed with 20–40 column volumes (CV) of Wash Buffer A (50 mM HEPES pH 7.5, 500 mM NaCl, 10% glycerol, 1% NP-40, 10 mM imidazole, 2.5 mM β-mercaptoethanol) in batch. Resin was centrifuged as above, resuspended in 6 CV of wash buffer A and poured into a column and washed with an additional 3 CV of wash buffer. Protein was eluted in 8 CV of Elution Buffer (50 mM HEPES pH 7.5, 500 mM NaCl, 10% glycerol, 1% NP-40, 200 mM imidazole, 2.5 mM β-mercaptoethanol; for FLAG-STREP tagged proteins, this buffer also contained 2.5 mM desthiobiotin). Eluted protein was passed twice over an amylose resin column (0.5 ml column resin per 1 cell culture). The resin was rinsed twice with 10CV of wash buffer B (50 mM HEPES pH 7.5, 150 mM NaCl, 10% glycerol, 5 mM DTT, 1 mM EDTA) and eluted in CTD buffer (50 mM HEPES pH 7.5, 150 mM NaCl, 10% glycerol, 5 mM DTT, 10 mM maltose, 1 mM EDTA)

Expression and purification of D. melanogaster P-TEFb. 600 ml of S9 cells were grown in suspension in baffled flasks at 27 °C in SF-900 II SFM media (Thermo Fisher Scientific), diluted to 1.5 million cells per ml, and infected with 1/10 culture volume of P-TEFb virus (a gift from J.T. Lis) and incubated at 27 °C with shaking at 75 r.p.m. 72 hrs after infection, the cells were collected at 500 g for 5 min and resuspended in P-TEFb binding buffer (50 mM HEPES pH 7.5, 500 mM NaCl, 10% glycerol, 1% IGEPAK, 2.5 mM Imidazole and 2.5 mM β-mercaptoethanol and protease inhibitors) and lysed in a dounce homogenizer. Lysate was centrifuged at 100,000 g for 30 min at 4 °C. Clarified supernatant was bound in batch to 0.5 ml TALON resin for 30 min at 4 °C, resin was pelleted at 250 g for 5 min with wash buffer, pelleted as before, washed and decanted into a gravity column and protein was eluted in 500 mM NaCl, 50 mM HEPES pH 7.5, 10% glycerol, 1% NP-40, 2.5 mM β-mercaptoethanol and 200 mM imidazole.

Phosphorylation and purification of MBP-CTD. MBP-D.melCTD6HIS was purified as described above, concentrated with a Vivaspin 30,000 MWCO concentrator (GE Life Sciences) to purified as described above, concentrated with a Vivaspin 30,000 MWCO concentrator (GE Life Sciences) to 10 CV of wash buffer B (50 mM HEPES pH 7.5, 10 mM MgCl2, 10 mM ATP and P-TEFb at 24 °C for ~48 h with additions of 1/10th reaction volume 100 mM ATP and 50 mM MgCl2 every 6–12 h). Elution was monitored by UV absorbance and the identity of the protein in each peak was verified by SDS–PAGE. All searches included phosphorylation of serine, threonine, and tyrosine as a variable modification and carbamidomethylation of cysteine as a fixed modification. Mass tolerances of 10 ppm for precursor and 0.02 Da were used for MS1 and MS2 peaks, respectively. Product ions typically observed during collisional activation (a, b, and y-type) were considered for spectrum matching. A fixed value PSMS validator filtered matches based on a maximum delta Cn of 0.05. For both PSMs and peptides, strict and relaxed target FDR settings were 0.01 and 0.05, respectively. PtmRS operating in PhosphoRS mode was used for phosphorylation site localization in which identified sites with isoform confidence probability less than 99% were further inspected manually.

Native gel electrophoresis. MBP-CTD fusions were diluted in 2X CN-PAGE buffer (25% glycerol, 62.5 mM Tris–HCl pH 6.8, 1% bromophenol blue) loaded onto 6 clear native-PAGE gels (SDS–PAGE gels with SDS omitted and a 4% acrylamide/bisacrylamide ratio of 48:1) and electroplorosed in running buffer (25 mM Tris–HCl, 192 mM glycine).

Size exclusion chromatography. A 100 μl protein standard mix containing thyroglobulin, apoferritin, aldolase, ovalbumin, and RNaseA, with stokes radii of 85, 67.1, 48.1, 30.5 and 16.4 kDa, respectively or MBP or MBP-CTD6HIS fusion proteins, were loaded on a Superose 6 10/300 column at 0.2 ml min⁻¹ in CTD buffer at 4 °C. Elution was monitored by UV absorbance (280 nm) and the elution volume (Ve) was void volume as determined by the elution of blue dextran and Ve is the total column volume.

Small angle X-ray scattering. SAXS data (I(q)) versus q, where q = 4πsin(q)/λ, were collected at the G1 beamline at the Cornell High Energy Synchrotron Source with an X-ray beam energy of 9.94 keV using dual Pilatus 100K-S detectors. Samples were exposed for 20 × 1 s frames. For each sample and at every concentration, buffer blanks were collected immediately before and after sample data acquisition. Data acquisition, buffer subtraction, data reduction, Guinier analysis to a qR<1.1, and molecular weight determination by extrapolation to I(0) and comparison to a glucose isomerase control were carried out using BioXtas RAW20 and plotted in Matlab. Buffer subtracted averaged scattering curves collected at three concentrations were averaged using PRIMUS, and distance distribution functions, Rg and Dmax were calculated using AUTOGNOM. Graphical scattering data for MBP was generated using CRYSTOL from 1ANF.pdb and plotted using AUTOGNOM. The Flory equation used to predict the Rg of the denatured CTD is, Rg = RgNβ, where Rg = 1.927, v = 0.598 (ref. 49) and N = 318, the number of amino acids in the D.melCTD portion of the MBP-D.melCTD fusion, including the C-terminal hexahistidine tag. Ensemble simulation models 4 generated using EOM 2.1, with averaged scattering curves and IANPdb as inputs34,35. Best agreement of Rg for EOM derived ensembles and the experimentally obtained data were obtained using a compact chain and native-like chain options for the apo and phospho CTD, respectively. Final ensembles were visualized using PyMoL, oriented by MBP, and displayed and aligned to the structure of RNA polymerase II from PDB 1T1W pdb for comparison26. Flexibility analysis was carried out using SCatter version 2.3 (ref. 43) and plotted in Excel or RAW.

Limited proteolysis. MBP-D.melCTDFLAG-STREP or MBP-H.sapCTD6HIS were end labelled on the terminal CTD repeat with casein kinase II (New England)
Biolab) in protein kinase buffer (50 mM Tris-HCl pH 7.5, 10 mM MgCl₂, 0.1 mM EDTA, 2 mM DTT, 0.01% Brij 35), 0.5 mM ATP and 0.33 mM AMPPNP (MP Biomedical) for 30 min at 30°C. After extract collection, BioRad BioSpin column. To reduce background signal from auto-labelled CKII in the MBP-H5-SPCTD6HIS experiments, CKII was first pre-incubated in kinase buffer with 4 mM ATP, in the absence of 32P-ATP, for three h at 30°C. This pre-incubition reaction was then added to a MBP-H5-SPCTD6HIS labelling reaction resulting in final reaction conditions as described above for labelled MBP-DmectFLAG-STR. End labelled CTD was digested in CTD buffer for 30 s at room temperature by adding protease concentrations ranging from 2 ng to 6,000 ng per 12 μl reaction. Reactions were quenched by the addition of PMSF to a final concentration of 10 mM. Mock reactions were performed with identical composition but without P-TEFb. The P-TEFb and mock reactions were mixed and immediately subject to limited proteolysis as described above, and fragments were resolved on a 12% SDS-PAGE gel. Signal intensity of the full length proteins were quantified using ImageJ and data from three experimental replicates, representing two separate labelling and kinase reactions, were plotted as the percentage of remaining undigested signal relative to the no protease control at each protein concentration.

Drosophila CTD sequence alignments. Amino acid sequences of the Pol II CTD for twelve Drosophila species, beginning with the sequence of YSPTSPNSTAS were aligned using MUSCLE28 improved manually, and visualized using BOXSHADE.

Fly strains and lethality test. UAS-Rpb1i and yw; Act-GAL4/Cyo and Cs-Gal4 fly lines were obtained from the Bloomington Stock Center (BDSC 36830, 4414, and 30839, respectively). Sequences encoding the Rpb1WT, Rpb1WT or Rpb1hi1, double FLAG-tags at the C-termini were subcloned into the pUASt-attB vector, followed by transformation into the genetic background C176 (for wing expression) to male flies by mating virgin female yw; Act-GAL4/Cyo/+ with double FLAG-tags at the C-termini were subcloned into the pUASt-attB vector, followed by transformation into the genetic background C176 (for ubiquitous expression) to male flies by mating virgin female yw; Act-GAL4/Cyo/+; UAS-Rpb1i, UAS-Rpb1WT or mut-CyO, CyO heterozygous females. Flies were raised at 25°C under 40% humidity and male adult offspring were identified and counted from the corresponding author on request. The data that support the findings of this study are available from the corresponding author on request.

References
1. Chapman, R. D., Heidemann, M., Hintermair, C. & Eick, D. Molecular evolution of the RNA polymerase II CTD. Trends Genet. 24, 289–296 (2008).
2. Corden, J. L. RNA polymerase II C-terminal domain: Tethering transcription to transcript and template. Chem. Rev. 113, 8423–8455 (2013).
3. Eick, D. & Geyer, M. The RNA polymerase II carboxy-terminal domain (CTD) for protein structure prediction. J. Mol. Biol. 296, 1111–1127 (2000).
4. Meinhart, A. & Cramer, P. Recognition of RNA polymerase II carboxy-terminal domain by 3'-RNA-processing factors. Nature 430, 223–226 (2004).
5. Cagan, R. D. & Corden, J. L. Structural studies of a synthetic peptide derived from the carboxy-terminal domain of RNA polymerase II. Proteins 21, 149–160 (1995).
6. Kumaki, Y., Matsuhashima, N., Yoshida, H., Nitta, K. & Hibiki, K. Structure of the YSPTSPS repeat containing two SPXX motifs in the CTD of RNA polymerase II. NMR studies of cyclic model peptides reveal that the SPTS turn is more stable than SPYS in water. Biochem. Biophys. Acta 1548, 81–93 (2001).
7. Biankiewicz, E. A., Moon Woody, A. & Woody, R. W. Conformation of the RNA polymerase II C-terminal domain: circular dichroism of long and short fragments. J. Mol. Biol. 297, 119–133 (2000).
8. Harding, M. M. NMRI studies on YSPTSPS: implications for the design of DNA bisintercalators. J. Med. Chem. 35, 4658–4666 (1992).
9. Noble, C. G. et al. Key features of the interaction between Pcf11 CTD and RNA polymerase II CTD. Nat. Struct. Mol. Biol. 12, 141–151 (2005).
10. Yang, C. & Stiller, J. W. Evolutionary diversity and taxon-specific modifications of the RNA polymerase II C-terminal domain. Proc. Natl Acad. Sci. USA 111, 5920–5925 (2014).
11. Corden, J. L., Eick, D. & Geyer, M. The RNA polymerase II CTD. J. Mol. Biol. 258, 413–419 (1996).
12. Zhang, J. & Corden, J. L. Phosphorylation causes a conformational change in the carboxy-terminal domain of the mouse RNA polymerase II largest subunit. J. Biol. Chem. 266, 2297–2302 (1991).
13. Laybourn, P. J. & Dahmus, M. E. Transcription-dependent structural changes in the C-terminal domain of mammalian RNA polymerase subunit IIa/o. J. Biol. Chem. 264, 6608–6616 (1989).
14. Echeverria, I. & Papoian, G. A. Structural heterogeneity and dynamics of the unfolded ensemble. Isr. J. Chem. 54, 1293–1301 (2014).
15. Uversky, V. N. Native unfolded proteins: a point where biology waits for physics. Protein Sci. 11, 739–756 (2002).
16. Meier, S., Blackledge, M. & Grzesiek, S. Conformational distributions of unfolded polypeptides from novel NMR techniques. J. Chem. Phys. 128, 052504 (2008).
17. Das, R. K. & Pappu, R. V. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. Proc. Natl Acad. Sci. USA 100, 13392–13397 (2003).
18. Groško, E. G., Hon, G., Chaconia, J. & Uversky, V. N. Biophysical constraints for protein structure prediction. J. Proteome Res. 2, 37–42 (2003).
23. Uversky, V. N. Size-exclusion chromatography in structural analysis of intrinsically disordered proteins. Methods Mol. Biol. 896, 179–194 (2012).
24. Brzustowicz, L. M., & Greenleaf, A. L. Functional studies of the carboxy-terminal repeat domain of Dro sophila RNA polymerase II in vivo. Genetics 140, 599–613 (1995).
25. Bernadó, P., & Svergun, D. I. Analysis of intrinsically disordered proteins by small-angle X-ray scattering. Methods Mol. Biol. 896, 107–122 (2012).
26. Czudnochowski, N., Böken, C. A. & Geyer, M. Serine-7 but not serine-5 phosphorylation primes RNA polymerase II CTD for P-TEFb recognition. Nat. Commun. 3, 842 (2012).
27. Uversky, V. N. Size-exclusion chromatography in structural analysis of intrinsically disordered proteins. Methods Mol. Biol. 896, 107–122 (2012).
28. Schüller, R. et al. Id2: a novel protein associated with boundaries of transcriptionally active chromatin. Genes Dev. 5, 1611–1621 (1991).
29. Njiru, S. S. et al. BioXTAS RAW, a software program for high-throughput automated small-angle X-ray scattering data reduction and preliminary analysis. J. Appl. Crystallogr. 42, 959–964 (2009).
30. Skou, S., Gillilan, R. E. & Ando, N. Synchrotron-based small-angle X-ray scattering of proteins in solution. Nat. Protoc. 9, 1727–1739 (2014).
31. Petoukhov, M. V. et al. New developments in the ATSAS program package for small-angle scattering data analysis. J. Appl. Crystallogr. 45, 342–350 (2012).
32. Receveur-Brechot, V. & Durand, D. How random are intrinsically disordered proteins? A small angle scattering perspective. Curr. Protein Pept. Sci. 13, 55–75 (2012).
33. Bernadó, P., & Svergun, D. I. Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering. Mol. Biophys. 8, 151–167 (2012).
34. Tria, G., Mertens, H. D. T., Kachala, M. & Svergun, D. I. Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering. IUCrJ 2, 207–215 (2015).
35. Quiocio, F. A., Spurlino, J. C. & Rodseth, L. E. Extensive features of tight oligosaccharide binding revealed in high-resolution structures of the maltodextrin transport/chemosensory receptor. Structure 5, 997–1015 (1997).
36. Kettenberger, H., Armache, K.-J. & Cramer, P. Complete RNA polymerase II elongation complex structure and its interactions with NTP and TFIIIS. Mol. Cell 16, 955–965 (2004).
37. Fontana, A., de Laureto, P. P., Spolaore, B. & Frare, E. Identifying disordered regions in proteins by limited proteolysis. Methods Mol. Biol. 896, 297–318 (2012).
38. Fontana, A. et al. Probing protein structure by limited proteolysis. Acta Biochim. Pol. 51, 299–321 (2004).
39. Fontana, A., Polverino de Laureto, P., Spolaore, B., Frare, E. & Zambonin, M. in Instrumental Analysis of Intrinsically Disordered Proteins 569–605 (John Wiley & Sons, Inc., 2010).
40. Wilkins, M. R. et al. Protein identification and analysis tools in the ExPaSy server. Methods Mol. Biol. 112, 531–552 (1999).
41. Hitz, W., Wiegert, U. & Adamietz, P. Stimulation of Proteinase K action by denaturing agents: application to the isolation of nucleic acids and the degradation of ‘Masked’ proteins. Eur. J. Biochem. 56, 103–108 (1975).
42. Brounawitz, M., Seneer, D. F., Shear, M. A. & Ackers, G. K. Quantitative DNase footprint titration: a method for studying protein-DNA interactions. Methods Enzymol. 130, 132–181 (1986).
43. Rambo, R. P. & Tainer, J. A. Characterizing flexible and intrinsically unstructured biological macromolecules by SAS using the Porod-Debye law. Biopolymers 95, 559–571 (2011).
44. Barron-Casella, E. & Corden, J. L. Conservation of the mammalian RNA polymerase II largest-subunit C-terminal domain. J. Mol. Evol. 35, 405–410 (1992).
45. Peter, A. et al. Mapping and identification of essential gene functions on the X chromosome of Drosophila. EMBO Rep. 3, 34–38 (2002).
46. Borg, M. et al. Polyelectrostatic interactions of disordered ligands suggest a physical basis for ultrasensitivity. Proc. Natl Acad. Sci. USA 104, 9650–9655 (2007).
47. Lenz, P. & Swain, P. S. An entropic mechanism to generate highly cooperative and specific binding from protein phosphorylations. Curr. Biol. 16, 2150–2155 (2006).
48. Konarev, P. V. et al. PRIMUS: a Windows PC-based system for small-angle scattering data analysis. J. Appl. Crystallogr. 36, 1277–1282 (2003).
49. Kohn, J. E. et al. Random-coil behavior and the dimensions of chemically unfolded proteins. Proc. Natl Acad. Sci. USA 101, 12491–12496 (2004).
50. Armache, K.-J., Kettenberger, H. & Cramer, P. Architecture of initiation–competent 12-subunit RNA polymerase II. Proc. Natl Acad. Sci. USA 100, 6964–6968 (2003).
51. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797 (2004).
52. Bischof, J., Maeda, R. K., Hediger, M., Karch, F. & Basler, K. An optimized transgenec system for Drosophila using germ-line-specific phiC31 integrases. Proc. Natl Acad. Sci. USA 104, 3312–3317 (2007).
53. Galloway, D. T., Frasch, M., Sauverber, H. & Liu, J. T. Characterization of a Drosophila protein associated with boundaries of transcriptionally active chromatin. Genes Dev. 5, 1611–1621 (1991).
54. Roote, J. & Prokop, A. How to design a genetic mating scheme: a basic training package for Drosophila genetics. G3: Genes, Genomes, Genetics 3, 353–358 (2013).

Acknowledgements We thank Roderick Acevedo for help with SAXS data collection, Richard Gillilan for his help at CHESS and John T. Lis for contributing the P-TEFb virus. We thank Xiangyun Qiu for assistance with SAXS data analysis. This work is based on research conducted at the Cornell High Energy Synchrotron Source (CHESS), which is supported by the National Science Foundation. We thank the National Institutes of Health and National Institutes of General Medical Sciences under NS8 award DMR-096384, using the Macromolecular Diffraction at CHESS (MacCHESS) facility, which is supported by award GM-103485 from the National Institutes of Health, through its National Institute of General Medical Sciences. We thank the TRIP at Harvard Medical School (NIH/NIGMS R01-GM084947) for providing transgenic RNAi fly stocks used in this study. This work was supported by GM04777 to D.S.G., NSF grant MCB-1515974 to S.A.S. B.P. was supported in part by NIAID T32 AI074451. Mass spectrometry work is supported by grants from the National Institutes of Health (ROI GM04096 to Y.J.Z. and R21EB018391 to J.S.B.) and Welch Foundation (F-1787 to Y.J.Z. and F-1155 to J.S.B.). Funding from the University of Texas System for support of the University of Texas System Proteomics Core Facility Network is gratefully acknowledged.

Author contributions B.P. designed experiments, analysed data, performed the purifications, SEC, SAXS, limited proteolysis and immunofluorescence experiments, and wrote the manuscript. F.I. generated transgenic flies, designed and conducted RNAi rescue experiments, performed the western blot, and commented on the manuscript. E.B.G. and S.A.S. designed experiments, analysed data, and commented on the manuscript. J.E.M., M.R.M., Y.J.Z. and J.S.B. designed, carried out and analysed mass spectrometry experiments. D.S.G. designed experiments, performed RNAi rescue and complementation experiments, analysed data, and prepared the manuscript.

Additional information Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://www.nature.com/reprintsandpermissions/

How to cite this article: Portz, B. et al. Structural heterogeneity in the intrinsically disordered RNA polymerase II C-terminal domain. Nat. Commun. 8, 15231 doi: 10.1038/ncomms15231 (2017).