Tampere University

FRANCESCO TABARO

Analysis of Chromatin and Proteins in Cancer

Tampere University Dissertations 354

Tampere University Dissertations 354

FRANCESCO TABARO

Analysis of Chromatin and Proteins in Cancer

ACADEMIC DISSERTATION
To be presented, with the permission of
the Faculty of Medicine and Health Technology
of Tampere University,
for public discussion at Tampere University
on 27th of November 2020, at 12 o'clock.

ACADEMIC DISSERTATION
Tampere University, Faculty of Medicine and Health Technology
Finland

Responsible supervisor and Custos	Professor Matti Nykter Tampere University Finland	
Pre-examiners	Professor Olli-Pekka Smolander Tallinn University of Technology Estonia	Associate professor Emidio Capriotti University of Bologna Italy
Opponent	Docent Christophe Roos University of Helsinki Finland	

Cover design: Roihu Inc.

ISBN 978-952-03-1794-2 (print)
ISBN 978-952-03-1795-9 (pdf)
ISSN 2489-9860 (print)
ISSN 2490-0028 (pdf)
http://urn.fi/URN:ISBN:978-952-03-1795-9

PunaMusta Oy – Yliopistopaino
Vantaa 2020

ACKNOWLEDGEMENTS

The works presented in this thesis have been carried out at the Faculty of Medicine and Health Technology at Tampere Universtity and the Department of Biomedical Sciences at Università degli Studi di Padova in Italy. I would like to sincerely thank Professor Matti Nykter for the professional and human support he provided over the years, for teaching and guiding me during this journey. I learned a lot from you. With your positive and sympathetic attitude you had a central role in helping me to reach my goals, even in the hardest moments. Working with you has been a great ride, thanks for giving me the opportunity. Also, I would like to thank Professor Silvio Tosatto and his collaborators Associate Professor Damiano Piovesan and Associate Professor Giovanni Minervini from the University degli Studi di Padova for letting me visit their laboratory and work in close collaboration with other researchers on critical and challenging projects. I feel honored to have had the chance to connect with all of you. I would also thank the Doctorate School of the Medicine and Health Technology Faculty for funding my last two years and half.

Thanks also to Professor Olli-Pekka Smolander and Professor Emidio Capriotti for reviewing this thesis. I think that with your advises and insights, the overall quality of the text and figures improved a lot. Thanks for caring and taking the time to carefully go through the book.

I would also like to thank the members of my steering committee, Professor Olli Yli-Harjia and the Dean of the Faculty Professor Tapio Visakorpi for being very welcome, kind and supportive during every single meeting we had. Thanks, it was an honor to report to you.

A special thanks to the people that helped me to translate the abstract in Finnish. You actually did all the work, Anssi, Kirsi, Matti, Juuso.

Next, I would like to thank all the people I have had the chance to meet during my stay in Finland: all the current and former members of the Nykter Lab, the people at Genevia Oy, colleagues from other research groups and friends, apologies for not

listing all your names. You guys were the fuel that kept me going. Thanks for being always positive, for the small talks, saunas, nights out and whatnot.

Last but not least, I would like to thank my family and my dear Gioia for the incredible love you keep showing and for letting me chase my dreams even when they sound crazy.

Bologna, November 2020
Francesco Tabaro

ABSTRACT

Gene expression is a thoroughly regulated process. The cooperation between proximal and/or distal regulative genomic elements allows precise positioning of the transcription machinery on gene's promoter and modulates the synthesis of transcripts. Transcription factors (TFs) are proteins able to bind these regulative loci. The availability of these sites is in turn regulated by chromatin structure. In cancer the delicate equilibrium between accessible and precluded TF binding sites gets altered. In prostate cancer (PCa), androgen stimulation plays a central role in sustaining cancer growth. Primary PCa, after treatment, recurs in about a third of cases with a more aggressive, androgen insensitive phenotype. Specific genetic alterations have been reported to drive primary cancer development and the transition to castration resistant prostate cancer (CRPC). From these notions, the connection between chromatin state, gene expression and PCa development can be hypothesized. The assay for transposase-accessible chromatin coupled with sequencing (ATAC-seq) was used to study the chromatin organization of samples representing different PCa progression stage collected at the Tampere University Hospital. This dataset was analyzed together with previously generated transcriptomic and publicly available chromatin immunoprecipitation followed by sequencing (ChIP-seq) data. From ATAC-seq data, peaks and differentially accessible regions (DARs) were detected. Correlation between ATAC-seq features and gene expression was calculated to assign each gene to a proximal or distal regulative region. At a global level, this analysis reported weak correlation between the two measurements. Nevertheless, expression of differentially expressed genes (DEG) showed a stronger correlation with accessible features. This observation supports the idea of alternative binding pattern utilization across PCa progression. To understand which transcriptional programs are involved in this process, TF binding sites were searched in candidate regulatory regions using ChIP-seq peaks. The transcription factor with highest number of binding sites across all ATAC-seq features is the androgen receptor (AR). Moreover, FOXA1 and HOXB13

were observed to co-localize with AR in two distinct sets of DARs with increased accessibility in PC or reduced accessibility in CRPC. This observation supports the idea of AR central role in driving PCa and lead to ask which TF co-modulate its activity in CRPC. To investigate this aspect and identify clusters of TF sharing target genes, a regulative network was built. Hierarchical clustering yielded two components: first a core, heavily connected module composed of AR, ERG, FOXA1 and ESR1, second a group of 43 TF sharing less target genes. This result confirms the central role of AR and highlights other TF, e.g. SP1, FLI1 and TP63 as its co-modulators.

All the identified TF share a fundamental structural organization: all of them have a DNA-binding domain and at least one regulatory domain. Moreover, the molecular structure of all these proteins show at least one intrinsically disordered region (IDR). These regions are flexible, display reduced hydrophobicity and net charge along their surface. In solution, intrinsically disordered proteins (IDPs) exist as a continuum of conformers with a structure that fluctuates from random coil to folded. To collect and organize literature-derived evidences of this phenomenon, the DisProt database was developed in 2006. Unfortunately, its updates were discontinued in 2013. To lead its manual annotation process, a dedicated web-service was created together with a completely re-designed web-application. While DisProt data is of the highest quality, the database size is limited. To extend intrinsic protein disorder annotation to the whole protein universe, MobiDB was created. This database collects data from eleven specialized external data sources and fifteen different tools for ID, secondary structure and low-complexity regions prediction. Using data from these resources the structure of above mentioned TFs was characterized and the emergent pattern of DNA-binding domain and IDRs detected.

Altogether these results demonstrate how integrated data analysis of multiple high throughput sequencing (HTS) measurements can help in dissecting the regulatory complexity of PCa by identifying sets of TFs involved cancer progression. Moreover, by utilizing these computational resources, structural features of identified proteins can be inferred. In general, these results provide a clear overview of the complexity of cellular phenomena, showcasing a data-driven workflow for detection of TFs involved in a disease and their structural characterization.

TIIVISTELMÄ

Geenien ilmentyminen on vahvasti säädelty biologinen prosessi. Proksimaalisten ja distaalisten säätelyalueiden yhteistyö mahdollistaa transkriptiokoneiston tarkan kohdentamisen geenin promoottoriin ja siten transkriptioaktiivisuuden säätelyn. Transkriptiotekijät sitoutuvat geenien säätelyalueille, joiden saavutettavuutta säädellään kromatiinirakenteen avulla, sitä avaamalla tai sulkemalla. Hienovarainen tasapaino saavutettavien ja suljettujen säätelyalueiden välillä muuttuu merkittävästi syöpäsoluissa. Eturauhassyövässä androgeenillä on keskeinen rooli jatkuvan syöpäkasvun ylläpitämisessä, ja se on myös yleinen hoitokohde. Hoitojen seurauksena noin kolmasosa eturauhassyöpäkasvaimista kehittyy aggressiivisiksi, androgeenista riippumattomiksi kasvaimiksi, joita kutsutaan yleisesti nimellä kastrattioresistentti eturauhassyöpä (castration resistant prostate cancer, CRPC). Tiettyjen geneettisten muutosten tiedetään johtavan eturauhassyövän tai sen kastraatioresistentin muodon kehittymiseen. Näistä lähtökohdista voidaan olettaa, että kromatiinirakenteen, geenien ilmenemisen ja eurauhassyövän etenemisen välillä on yhteys . Väitöskirjassa kromatiinirakennetta tarkasteltiin ATAC-seq (transposase-accessible chromatin coupled with sequencing) - menetelmällä Tampereen yliopistollisessa sairaalassa kerätyistä potilaiden eturauhassyöpänäytteistä, jotka edustivat syövän eri vaiheita. Analyysissä hyödynnettiin aikaisemmin samoista näytteistä tuotettua geenien ilmenemisdataa (RNA-seq), sekä julkisesti saatavilla olevaa transkriptiotekijöiden sitoutumisdataa (ChIP-seq). ATAC-seq datan avulla tunnistimme useita syöpään liittyviä muutoksia kromatiinin rakenteessa. Yhdistämällä havaitut kromatiinirakenteen muutokset geenien ilmenemismuutoksiin pystyimme liittämään geenit säätelyalueisiinsa. Vaikka koko genomin mittakaavassa yhteydet säätelyalueiden ja geenien ilmentymistasojen välillä olivat heikkoja, syövän etenemiseen liittyvien geenien säätelyalueiden muutokset liittyivät selkeämmin niiden ilmenemiseen. Saadut tulokset tukevat ajatusta siitä, että eturauhassyövän etenemiselle on tunnusomaista transkriptiotekijöiden sitoutumiskohtien muuttuminen patologisella tavalla.

Ymmärtääksemme, mitkä transkriptiomekanismit liittyvät syövän kehittymiseen ja etenemiseen, kävimme läpi transkriptiotekijöiden sitoutumisalueita ChIP-seq datan perusteella. Androgeenireseptorilla (AR) oli suurin määrä sitoutumiskohtia ATAC-seq-analyysissa havaituilla muuttuneilla kromatiinialueilla. Lisäksi FOXA1 ja HOXB13 transkriptiotekijöiden havaittiin sitoutuvan samoihin kohtiin androgeenireseptorin kanssa alueilla, jotka avautuivat aikaisen vaiheen eturauhassyövissä ja sulkeutuivat CRPC:ssä. Saatu havainto tukee AR-geenin keskeistä roolia eturauhassyövän etenemisessä ja saa pohtimaan, mitkä transkriptiotekijät liittyvät sen aktiivisuuden muokkaamiseen CRPC:ssä. Vastataksemme tähän kysymykseen tunnistimme transkriptiotekijäjoukkoja, joilla on paljon yhteisiä kohdegeenejä. Transkriptiotekijöiden ryhmittely hierarkisen klusteroinnin avulla paljasti kaksi ryhmää: Ensimmäiseen ryhmään kuuluivat geenit AR, ERG, FOXA1 ja ESR1, jotka muodostavat AR-säätelyn perustan ja liittyvät vahvasti toisiinsa. Toiseen ryhmään kuului 43 transkriptiotekijää, joilla oli vähemmän yhteisiä kohdegeenejä. Saatu tulos validoi AR-geenin keskeistä roolia ja nostaa esiin muiden säätelijöiden, kuten SP1:n, FLI1:n ja TP63:n, merkityksen AR:n rinnakkaissäätelijöinä .

Kaikilla transkriptiotekijöillä on samankaltainen proteiinirakenne: ne sisältävät DNA-sitoutumisdomeenin ja vähintään yhden säätelyyn liittyvä domeenin eli proteiinin osa-alueen. Tämän lisäksi kaikilla on vähintään yksi rakenteellisesti järjestäytymätön domeeni. Nämä järjestäytymättömätömät alueet ovat taipuisia, vain lievästi hydrofobisia, eikä niillä tyypillisesti ole sähkövarausta. Järjestäymättömän proteiinirakenteen omaavilla proteiineilla (intrinsically disordered proteins, IDPs) on nesteessä useita mahdollisia rakenteita jotka voivat vaihdella satunnaisesta rihmasta täysin järjestäyneeksi , laskostuneeksi muodoksi. DisProt-tietokanta luotiin näitä proteiineja tutkivan kirjallisuuden kartoittamiseksi ja yhteenkokoamiseksi. Uusi verkkosivusto luotiin ohjaamaan julkaistun tiedon kuratointia ja tietokanta toteutettiin uutena verkkosovelluksena. Vaikka DisProt tietokannan data on huippulaatuista, sen sisältämän tiedon määrä on vielä rajallinen. MobiDB- tietokanta luotiin lisäksi, jotta järjestymättömien alueiden annotointi voidaan tehdä kattavasti kaikille tunnetuille proteiineille . ModiDB tietokantaan on kerätty tietoja yhdestätoista eri tietokannasta ja viisitoista eri algoritmia järjestäytymättömän proteiinirakenteen, proteiinin sekundäärirakenteen ja epätyypillisen aminohappokoostumuksen omaavien alueiden ennustamista varten. Näiden työkalujen avulla analysoimme edellämainittujen transkriptiotekijöiden DNA-sitoutumisdomeenien ja järjestymät-

tömien alueiden rakennetta. Nämä tulokset osoittavat kuinka eri tyyppisten uuden sukupolven sekvensointimenetelmien tulosten analysointi yhdessä auttaa selvittämään monimutkaisia säätelyprosesseja. Transkriptiotekijöiden analyysillä voidaan paremmin ymmärtää eturauhassyövän syntyä ja etenemistä kastraatioresistentiksi muodoksi. Lisäksi kehitettyjen menetelmien avulla pystytään selvittämään tunnistettujen transkriptiotekijöiden proteiiniakennetta. Väitöskirjassa saadut tulokset tarjoavat yleiskuvan solunsisäisten prosessien monimutkaisuudesta ja tuovat esiin laskennallisia lähestymistapoja tauteihin liittyvien transkriptiotekijöiden tunnistamiseksi ja karakterisoimiseksi.

CONTENTS

List of Figures

List of Tables

ABBREVIATIONS

API	application programming interface
AR	androgen receptor
ATAC-seq	assay for transposase-accessible chromatin coupled with sequencing
BMRB	BioMagResBank
bp	base pair
BPH	benign prostate hyperplasia
CATH	Class, Architecture, Topology and Homology database
ChIP-seq	chromatin immunoprecipitation followed by sequencing
CNA	copy number alteration
CRPC	castration resistant prostate cancer
DAR	differentially accessible region
DB	database
DBMS	database management system
DE	differential expression
DEG	differentially expressed genes
DIBS	Disordered Binding Site database
DNA	deoxyribonucleic acid
ELM	Eukaryotic Linear Motifs
FDR	false discovery rate
FELLS	Fast Estimator of Latent Local Structure

FESS	Fast Estimator of Secondary Structure
FuzDB	Fuzzy Complexes Database
GLM	generalized linear model
GTRD	Gene Transcription Regulation Database
GUI	graphical user interface
H3K27ac	Histone 3 Lysine 27 acetylation
H3K27me1	Histone 3 Lysine 27 mono-metylation
H3K27me3	Histone 3 Lysine 27 tri-metylation
H3K4me3	Histone 3 Lysine 4 tri-metylation
HGP	Human Genome Project
HOMER	Hypergeometric Optimization of Motif EnRichment
HTS	high throughput sequencing
ID	intrinsic protein disorder
IDEAL	Intrinsically Disordered proteins with Extensive Annotations and Literature
IDP	intrinsically disordered protein
IDR	intrinsically disordered region
JS	Javascript
JSON	Javascript Object Notation
LIP	linear intercting peptide
LOH	loss of heterozygosis
MACS	Model-based Analysis of ChIP-Seq
MFIB	Mutually Folding Induced by Binding database
MRI	magnetic resonance imaging
NGS	next-generation sequencing
NMR	nuclear magnetic resonance
PC	primary prostate cancer
PCa	prostate cancer

PDB	Protein Data Bank
PIC	pre initiation complex
PSA	prostate-specific antigen
RAM	random access memory
RCI	Random Coil Index
REST	Representational state transfer
RIN	residue interaction network
RNA	ribonucleic acid
RNA-seq	RNA-sequencing
RONN	Regional Order Neural Network
RP	radical prostatectomy
SIFTS	Structure Integration with Function, Taxonomy and Sequence
SVM	support vector machine
TAD	topologically associated domain
TBP	TATA binding protein
TCGA	The Cancer Genome Atlas
TF	transcription factor
TFBS	transcription factor binding sites
TMV	Tobacco mosaic virus
TSS	transcription start site
TURP	trans-uretral radical prostatectomy

ORIGINAL PUBLICATIONS

Publication I J. Uusi-Mäkelä*, E. Afyounian*, F. Tabaro*, T. Häkkinen*, A. Lussana, A. Shcherban, M. Annala, R. Nurminen, K. Kivinummi, T. L. Tammela, A. Urbanucci, L. Latonen, J. Kesseli, K. J. Granberg, T. Visakorpi and M. Nykter. Chromatin accessibility analysis uncovers regulatory element landscape in prostate cancer progression. *bioRxiv* (2020). DOI: 10.1101/2020.09.08.287268. eprint: https://www.biorxiv.org/content/early/2020/09/09/2020.09.08.287268.full.pdf. URL: https://www.biorxiv.org/content/early/2020/09/09/2020.09.08.287268.

Publication II D. Piovesan*, F. Tabaro*, I. Mičetić, M. Necci, F. Quaglia, C. J. Oldfield, M. C. Aspromonte, N. E. Davey, R. Davidović, Z. Dosztányi, A. Elofsson, A. Gasparini, A. Hatos, A. V. Kajava, L. Kalmar, E. Leonardi, T. Lazar, S. Macedo-Ribeiro, M. Macossay-Castillo, A. Meszaros, G. Minervini, N. Murvai, J. Pujols, D. B. Roche, E. Salladini, E. Schad, A. Schramm, B. Szabo, A. Tantos, F. Tonello, K. D. Tsirigos, N. Veljković, S. Ventura, W. Vranken, P. Warholm, V. N. Uversky, A. K. Dunker, S. Longhi, P. Tompa and S. C. E. Tosatto. DisProt 7.0: a major update of the database of disordered proteins. *Nucleic acids research* 45 (D1 Jan. 2017), D219–D227. ISSN: 1362-4962. DOI: 10.1093/nar/gkw1056. ppublish.

Publication III D. Piovesan*, F. Tabaro*, L. Paladin, M. Necci, I. Micetic, C. Camilloni, N. Davey, Z. Dosztányi, B. Mészáros, A. M. Monzon, G. Parisi, E. Schad, P. Sormanni, P. Tompa, M. Vendruscolo, W. F. Vranken and S. C. E. Tosatto. MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions

in proteins. *Nucleic acids research* 46 (D1 Jan. 2018), D471–D476. ISSN: 1362-4962. DOI: 10.1093/nar/gkx1071. ppublish.

* Equal contribution

Author's contribution

Publication I — Performed all gene expression-related analyses: differential gene expression and batch correction on RNA-seq, alignment, quantification and differential gene expression on smallRNA-seq. Participated in the design and performed integrated analysis with ATAC-seq dataset. Performed downstream integrated analysis with publicly available ChIP-seq data and ATAC-seq.

Publication II — Participated in the design of database schema and in the migration of data. Coordinated transition from previous database release and technologies to new non-relational platform. Designed and implemented REST-based back-end service and web-based front-end graphical user interface. Developed the biocurator interface and simple search system.

Publication III — Participated in the design of database schema. Implemented REST-based back-end service. Performed transition to new front-end technologies and developed web-based graphical user interface as well as sequence and structure viewers. Supervised development of features viewer. Designed and developed the search system related subsystems.

1 INTRODUCTION

Conditional activation of gene expression regulates intracellular concentration of transcripts. A gene is expressed if transcription factors (TFs) bind on its promoter, trigger the formation of a pre initiation complex (PIC) and the RNA polymerase II is able to leave the promoter and transcribe the entire gene body. This mechanism requires coordinated interaction of proximal and distal TF. Signal transduction pathways are cellular systems devoted to sense and transmit an extracellular signal to the nucleus and stimulate gene expression. The final effectors of these signal cascades are TFs.

In the nucleus of eukaryotic cells, genomic DNA interacts with specialized proteins to form chromatin whose basic discrete units are nucleosomes. Gene transcription requires precise chromatin structural organization. Its three-dimensional structure may block the interaction between TF and DNA while nucleosome positioning may cause RNA polymerase to stall. Cellular stimuli may result in chromatin structure reconfiguration allowing or inhibiting gene expression. The combination of chromatin structure, TFs intracellular concentration and reaction to external stimulation drives gene expression which is at the basis of every cellular process.

Alterations to these mechanisms lead to pathological phenotypes. In cancer, aberrant regulation of signaling pathways alters cellular phenotype, cell cycle and results in uncontrolled proliferation. Prostate cancer (PCa) develops from prostate epithelium. These cells are physiologically sensitive to testosterone stimulation that is required for development of primary and secondary male sexual traits in physiological condition. Androgen receptor (AR) is the intracellular sensor for testosterone. Testosterone-bound AR dimerizes and migrates to the nucleus where it binds androgen responsive elements and activates expression of AR-inducible genes. Upon upregulation, AR stimulation leads to AR-inducible genes overexpression, uncontrolled cellular proliferation and tumor mass formation. After first-line treatments, in about a third of cases, PCa recurs with a more aggressive, androgen-insensitive phenotype. This

observation leads to the hypothesis that altered gene expression and alternative utilization of regulative programs can be explained by diverse and extensive chromatin reconfiguration at different disease stages.

The AR is an example of intrinsically disordered protein (IDP). This protein class is characterized by a flexible tertiary structure, reduced hydrophobicity and net charge. Many proteins, especially in higher eukaryotes, display at least one intrinsically disordered region (IDR). As opposed to globular proteins, IDPs have no enzymatic activity but have an important role in molecular recognition processes such as protein-protein, protein-ligand and protein-DNA interactions. Their flexibility and adaptability allows a one-to-many interaction pattern. Recently, it has been shown that intrinsic protein disorder is involved in membrane-less organelles formation by phase separation of nuclear factors controlling gene expression [1] and the genome scanning performed by TFs to select binding sites [2]. Because of the central role in interaction networks, phase separation and aggregation, experiments and computational resources for analysis and annotation of IDP are widely available. DisProt [3] is a repository of manually curated annotations on intrinsic protein disorder. Manual curation ensures the highest data quality and allows the generation of a controlled vocabulary to describe the molecular aspects of these proteins. Development of curated and integrated data sources for the retrieval and visualization of IDP annotation is thus crucial for their study. Main limitation of this approach is its throughput. Because of this, prediction tools and indirect evidences from third-party sources have been collected in MobiDB. These two databases together provide complete and extensive structural and functional annotation of intrinsic protein disorder. Among the others, these tools simplify the structural characterization of TF involved in any disease including PCa, supporting planning of experimental study these proteins and with implications in drug and therapy design.

2 LITERATURE REVIEW

2.1 Epigenetic control of gene expression

Each somatic human cell, in its nucleus, contains forty-six molecules of genomic DNA accounting for six billion base pairs (bp). The length of the DNA molecule composing chromosomes varies from 85 mm to 16 mm with the longest one (chromosome 1) made of 250 Mbp and the smallest (chromosome 21) made of 60 Mbp. If connected, these molecules would be about 1.8 m long. The average diameter of a human somatic cell is 10 µm, and the cell nucleus has a diameter of 6 µm. These dimensions impose a spatial constraint on the nuclear organization of DNA molecules implying a compression mechanism to store the genetic information. Chromatin is the complex of DNA and proteins in the nucleus devoted to this task.

2.1.1 Chromatin organization

First and fundamental chromatin units are nucleosomes (Figure 2.1A). Each of them is formed by eight histonic subunits. Histones are basic proteins with a core structural domain highly conserved across all eukaryotic organisms. Four couples of subunits form one nucleosome: two copies of histone H2A, H2B, H3 and H4, respectively. Multiple histone variants have been identified and have been associated with different biological processes: utilization of histone variants marks genomic loci for specific process, e.g. H2A.Z and H3.3 variants have been associated with reduced nucleosome stability, nucleosome depleted regions at active genes promoters and transcription initiation. On the other hand, H2A.X is involved in DNA breakage repair and V(D)J recombination in lymphocytic cell differentiation. [4, 5]

Histones structure is characterized by the histonic domain fold and an unfolded N-terminal tail of about 30 residues. This long tail is recognized by epigenetic readers and

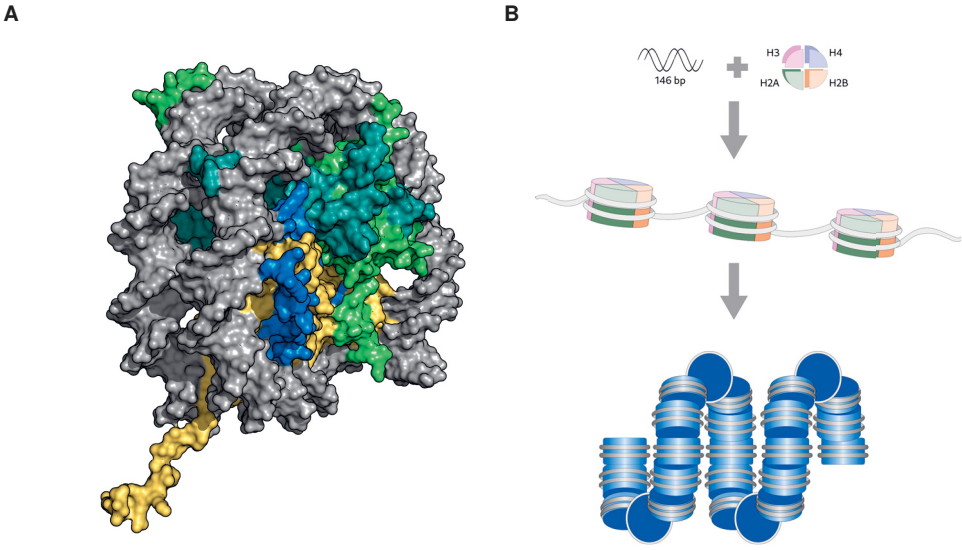


Figure 2.1 Nucleosome structure and basic levels of chromatin organization. **A.** Nucleosomes are composed of 8 histonic subunits. DNA wraps around histones with a period of 146 bp. Rendered from PDB structure 1AOI [6]. **B.** Histones and DNA interact to form nucleosomes. Multiple DNA-bound histones form a "bead on a string" structure. Nucleosomes interact to achieve denser organization forming the 30 nm fiber. Adapted from 7.

writers and is target of post translational reversible modifications, e.g. methylation, acetylation, phosphorylation, ubiquitination, sumoylation and lactylation [8, 9]. Different functional meaning have been assigned to modification of histone residues: H3K4me3 has been associated with transcriptional repression, while H3K27me1 and H3K27ac have been associated with active transcription [10, 11]. These unfolded regions are important regulators of chromatin structure and are implied in epigenetic control of gene expression.

DNA binds a nucleosome by wrapping around it (Figure 2.1A). Under optical microscope, the complex of nucleosomes and DNA looks like a "bead on a string": DNA wraps around the histonic octet with a periodicity of 146 bp and a short inter spread stretch of DNA separates each couple of nucleosomes [12]. Multiple nucleosomes may interact forming a fiber-like structure called 30 nm fiber (Figure 2.1B). Interactions among histone tails from adjacent nucleosomes reduce their spatial distance and histone H1 stabilizes the interaction forming this structure achieving 50-fold compression of the genetic information [5].

Other non histonic proteins have the ability to bind chromatin and induce even

higher order condensation. These proteins give rise to tertiary structures and form oligomeric molecular complexes by coordinating multiple chromatin fibers. For example, the Polycomb proteins, bind specific sequences on the genome and are responsible for deposition of H3K27me3, a repressive histone mark, and induce chromatin compactation [13]. The chromatin packing process achieves an extremely efficient degree of compression and limits the interaction between transcription factors, transcription machinery and their genomic targets.

From an evolutionary perspective, chromatin may have evolved primarily as a mechanism to repress gene expression, viral insertions and transposition events in eukaryotic genomes [14]. Regions of active and inactive transcription have been detected in nuclei of eukaryotic cells via chromatin conformation capture experiments [15]. Active compartments are associated with euchromatic nuclear regions, loosely packed DNA and higher gene expression. On the other hand, inactive compartments are associated with heterochromatic regions, denser chromatin and reduced gene expression. [16] Chromatin gets remodeled as a response to external stimuli, e.g. in macrophages the TLR pathway activates NF-κB sensitive genes. Here, two waves of expressed genes can be detected, with the latter being induced by the products of the former. [17]

2.1.2 Interplay between chromatin structure, transcription factors and gene expression

Chromatin organization represents a fundamental layer of gene expression regulation: by chromosomal packing the access to genetic information is denied to the transcription machinery and thus, the synthesis of genic products inhibited. TFs, on the other hand, are proteins responsible for activation of gene expression at specific genomic loci. Basal TFs recognize DNA sequences located in proximity of genes transcription start site (TSS), bind them and induce formation of pre initiation complex (PIC). Nevertheless, TFs have also distal binding sites. Binding to these elements has been shown to be required for releasing the transcription machinery from the promoter. In mammals, enhancers dysfunction is linked to developmental malformations highlighting their central role in coordinating transcription [20, 21]. Transcription factors bound to proximal elements are required for effective assembly of the transcription machinery but enhancer binding and activation influences transcription rate [22].

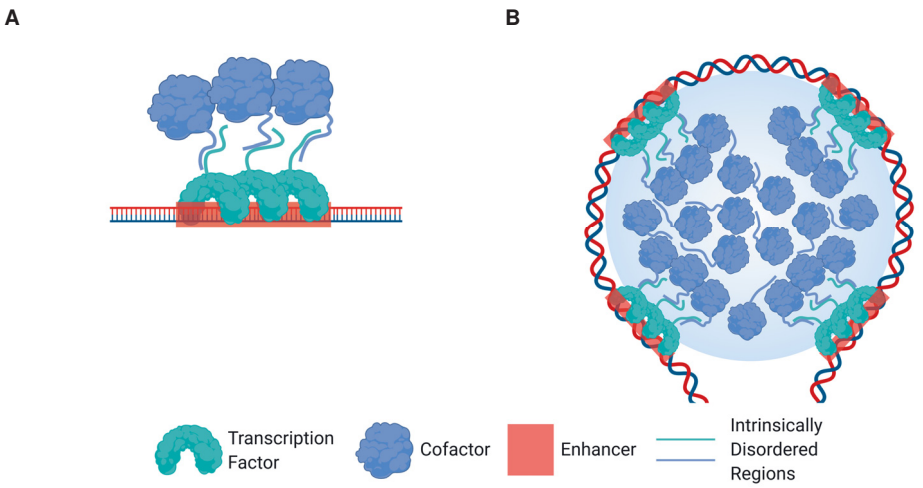


Figure 2.2 Models of interaction between TF and transcriptional coactivators through IDR. **A.** At enhancer loci, IDRs mediate interaction between TFs bound to DNA and soluble transcriptional coactivators. **B.** At super-enhancer loci, enhancer-bound TFs interact with multiple copies of coactivator proteins forming phase-separated droplets. Inside these droplets, interaction among TF, RNA polymerase subunits and other cofactors is facilitated. Adapted from 19.

Transcription factors bound to enhancer elements can recruit chromatin remodellers. These proteins induce structural changes in chromatin conformation resulting in a loop that puts the TF in spatial proximity of the PIC assembling on a gene promoter. The interaction between basal TF and enhancer-bound TF results in the release of DNA polymerase from promoter and initiation of transcription. Recently, gene expression has been associated with the idea of transcription factories [1]. These, are loci of stable enhancer-promoter interaction, polymerase condensation and transcript initiation. Moreover, in some other cases, the enhancer-promoter interaction has been observed to persist during transcription elongation phase [23]. On top of this, enhancers can show additive effect: more than one enhancer can contact a gene promoter resulting in increased transcript synthesis [22]. Some genomic loci longer than regular enhancers display unusual enrichment for TF binding sites and H3K27ac histone modifications. They have been shown to work as interaction hubs and to be implied in regulation of multiple genes; because of this they have been termed super-enhancers [24].

Enhancer cis-regulatory function does not extend throughout entire chromosomes, in fact it is bound within topologically associated domains (TADs). These are genomic compartments with preferential intra-domain interactions. A TADs

forms between two distal convergent CTCF binding sites, defined insulators. The CTCF transcription factor binds CCCTC sequence motifs on the genomic DNA and recruits Cohesin monomers. Between two convergent CTCF binding sites, cohesin can dimerize forming a ring around DNA resulting in an extended loop defined TAD [25, 26]. Genes located in the same domain are regulated similarly: [27] it has been shown that enhancers interact preferentially with promoters within the same TAD [28, 29, 30]. Notably, transcription stimulation has been shown to correlate with an augmented number of enhancer-promoter interactions [28, 30]. However, albeit having been postulated to be the fundamental unit of gene expression, TAD cannot fully explain observed gene expression variability [31].

Transcription factors regulate gene expression by binding to target sequences on the genome. These are recognized by specialized structural domains called DNA binding domains. However, TFs activity is influenced by a number factors, ranging from local chromatin structure and post-translational modifications, DNA methylation and others [32, 33]. Moreover, empirical observations show that among all possible binding sites available in the genome, only a subset is occupied *in vivo*. Different mechanisms have been proposed to explain this observation, either involving cooperative binding of multiple TFs [34] or the sequence composition in the vicinity of binding domain [35].

2.2 Prostate cancer

2.2.1 Epidemiology, diagnosis and clinical treatment

In 2020, in the US, prostate cancer (PCa) will be the most newly diagnosed cancer, accounting for more than thirty thousand deaths (10% of total cancer deaths) [36]. Prostate cancer is described as an age-related disease: the probability of developing it doubles from 60 to 70 years and men older than 80 have more than 10% chance of disease development [36]. Big geographical and ethnic variations in diagnosis rates exist. These differences are partly due to different practices in prophylactic screenings, lifestyle and migration patterns. In the last 40 years a general increase in diagnosis has been observed and it has been correlated with the increased utilization of prophylactic screening [37, 38]. Along with age and African ancestry, PCa risk factors include obesity, smoking and stature. Familial history and a susceptible genetic background

are considered risk factors as well [38].

Prophylactic screening are based on detection and quantification of prostate-specific antigen (PSA). This is a peptidase secreted by the prostatic epithelium that, in physiological conditions, liquefies semen and should not be detected in plasma. Its presence in blood is used as biomarker, and a quantification assay is routinely used in clinical practice. PSA blood concentration correlates with PCa grade and is used to stratify the risk of PCa development and its status. A threshold value of $10\,\mathrm{ng\,mL^{-1}}$ would entitle a patient for prostate biopsy. Prostate cancer diagnosis is based on microscopic evaluation of prostate tissue obtained by needle biopsy, the procedure implies a pathologist grading the sample with a Gleason score from 1 to 5 based on morphological characteristics of the tissue sample. Patient risk is then stratified using the PSA concentration, histological evaluation and clinical stage. To improve risk stratification MRI [39, 40, 41] and new biomarkers have been tested [42, 43]. An epigenetic test quantifies DNA methylation and reaches discriminatory power similar to PSA [43]. Recently, an automatic method for PCa detection from whole slide scan images using machine learning has been proposed [44] and genomic characterization from free circulating tumor DNA are either available commercially and or under active development in academic settings [45, 46, 47]. From tissue biopsy molecular biomarkes can be used to classify tumor aggressiveness and identify more aggressive cases.

The risk of dying from PCa depends on age and comorbidity. Today, the probability of dying from other causes is greater than the probability of dying from PCa. For all stages of PCa combined, the 5 years overall survival rate is 98% [36]. The 10 years risk of death ranges from 3% to 18%, while, for men with comorbidity, 10 years mortality rate from other causes rises to 33% or higher [48, 49]. Men diagnosed with localized disease have mainly treatment choices depending also on the detected prostate-specific antigen (PSA) levels: expectant management or hormonal therapies. The first option consists of watchful waiting, based on palliative cures of symptoms, and active surveillance. This option involves repeated PSA measurements and biopsies to monitor the disease progression. The other option represents the most effective alternatives for more severe clinical manifestations (e.g. those with PSA level greater than $10\,\mathrm{ng\,mL^{-1}}$) [49]. The main goal is to reduce testosterone production. Multiple strategies exist to achieve this: orchiectomy or surgical removal of testis is the most effective treatment able to deplete up to 95% of testosterone production. Medical

castration is another strategy consisting in the utilization of chemical compounds to inhibit testosterone secretion from testis.

Prostate cancer relapse happens in about a third of case, even after years [50, 51]. First-line treatment for these cases is androgen deprivation therapy. This therapy, although effective, has some adverse effects: it is associated with toxicity, decreased bone mineral density, metabolic change, sexual disfunction, hot flashes, cardic morbidity and cognifive disfunctions [49].

2.2.2 Genomics

Early studies The genetics and genomics of prostate cancer have been studied since the 80s: first identified mutations were large chromosomal alterations in chromosomes 10p, 10q, 8p, 8q, and 17q [52, 53, 54, 55]. These loci code for important oncogenes and tumor suppressor genes such as *TP53*, *RB1*, *NKX3-1* and *PTEN*. The first observed alterations were 10q24 deletion and mutations in 8p [52]. In 1994, loss of heterozygosis (LOH) was observed in chromosome 17p in a locus associated with expression of TP53 [56]. Later, in 1990, *RB1* deletion was reported to induce more aggressive phenotype in a cell line model [57]. In early 90s deletion of 8p was confirmed by multiple independent groups and, in 1997, the tumor suppressor gene *NKX3-1* identified in 8p21 [58, 59]. In 1992, the first *AR* mutation associated with primary PCa was reported [60]. In later time, *AR* mutations, especially amplifications have been associated with CRPC [61]. The *PTEN* tumor suppressor gene was identified in chromosome 10q23.1 in 1997 [62] and shown to be involved in downregulation of PI3K/Atk pathway. The first amplification of the *c-Myc* locus was observed in 1986 [63], and was consistently detected in many subsequent studies. Alterations of this oncogene have been associated with CRPC progression when co-occurrencing with *PTEN* mutations [64, 65].

Structural alterations Copy number alteration (CNA) are commonly detected in primary PCa. About three quarters of primary tumors display some kind of CNA [66, 67]. Common alterations are deletions localized in 8p, 13q, 6q, 16q, 18q and 9p. Common gains are observed in CRPC in chromosome 7, 8q and X [66]. Moreover, in about 50% of cases a fusion event between *TMPRSS2* and *ERG* is detected (TMPRSS2:ERG) [68]. This mutation puts the *ERG* gene, which codes for an ETS

transcription factor, under control of the *TMPRSS2* promoter, which is sensitive to androgen stimulation. This fusion achieves androgen-dependent transcriptional control of *ERG* resulting in enhanced cellular motility. Other members of the ETS gene family have been observed fused to *TMPRSS2*, e.g. *ETV1* [68], *ETV4* [69] and *FLI1* [70]. Deletion of chromosome 10q implies deletion of the *PTEN* tumor suppressor gene. This gene is involved in the PI3K/Akt pathway and clonal fraction of this mutation correlates with cancer progression. The TMPRSS2:ERG fusion has been identified as gatekeeper of PCa characterizing the transition from BPH to primary PCa. Accumulating mutations in aforementioned tumor suppressor genes and oncogenes coupled with treatment-induced clonal selection drive the transition to CRPC. Clonal fusion events have also been identified as happening in different locations within tumor nuclei generating multiple cellular subpopulations with convergent evolutionary trajectories [71].

Primary PCa In recent years, with the advent of next generation sequencing technologies, large cohorts have been analyzed confirming early genomic observations. The The Cancer Genome Atlas (TCGA) project characterized 333 primary PCa samples [72]. According to common genomic features, samples were clustered in six groups. The first four involve fusion or overexpression of ETS genes: the first and largest cluster with *ERG*, the second with *ETV1*, the third with *ETV4* and the fourth, smaller cluster with *FLI1*. In total 53% of samples showed a mutation involving an ETS gene. The remainder portion exhibit missense mutations in *SPOP*, *FOXA1* and *IDH1*. Commonly observed CNA involved amplification of chromosome 8, deletion of 6, 13 and 16 with different proportions across clusters. Although many structural variants have been identified and used to classify samples, PCa is generally described as a low-mutation cancer, and in general the overall mutational burden is lower than the burden showed by other tumors of epithelial origin. The DNA methylation pattern was found to be altered in these samples and widespread hypermethylation detected. Methylation-based clustering defines four clusters largely overlapping with the previously defined ones. The observed methylation patterns suggest widespread genomic silencing at early stage of the disease. As PCa growth is sustained by testosterone, a steroid hormone with a nuclear receptor coded by the *AR* gene, AR activity was quantified. *AR* is located on chromosome Xq12 in a frequently amplified locus. In the TCGA cohort, AR activity is increased in *SPOP1* and *FOXA1* clusters. On

the other hand, clusters showing ETS fusions do not show increased AR activity. Moreover, *SPOP1* mutations are mutually exclusive with TMPRSS2:ERG fusions and the transition to CRPC in presence of *SPOP1* mutation is driven by the accumulation of mutations in *PTEN* and *AR*.

Castration resistant prostate cancer Hallmark of the transition to CRPC are independence from androgen stimulation, loss of chromosome 17p, *TP53* and resistance to apoptosis due to *BCL2* overexpression [73], altered cell cycle due to mutations in *RB1* and CDK genes as well as abnormal activation of PI3K/Akt pathway and alterations to DNA repair system [74]. The earliest evidences of *AR* mutations were reported in the fist half of 90s with observations regarding CRPC cells growth with minimal testosterone stimulation or in its total absence [60, 61, 75, 76, 77, 78, 79, 80]. Sequencing experiments have identified recurrent structural alterations either in the *AR* gene body [81] or in an upstream enhancer region [82]. Nevertheless, other mechanisms may lead to AR reactivation: alterations in splicing producing AR variants lacking the ligand-biding domain or protein stabilization from other cofactors [83] are commonly observed. Moreover, paracrine hormonal stimulation has been reported and linked to gain of function mutations in enzymes of the dihydrotestosterone biosynthetic pathway [84]. Activation of PI3K/Akt pathway is due to deletions of the *PTEN* tumor suppressor gene in 50% of metastatic CRPC or, less frequently, amplification of other genes from the same pathway [85]. The *TP53* gene localizes in chromosome 17p, a frequently deleted locus, linked to cancer recurrence and metastasis [86, 87, 88, 89, 90, 91]. Overexpression of *BCL2* has been reported since 1993 [92, 93, 94, 95] when resistance to chemotherapy in cell lines was first observed [96]. Mutations in *BRCA1*, *BRCA2* and *ATM* are observed in about 20% of metastatic CRPC. Also the WNT pathway has been reported to be altered in about a fifth of cases, most frequently because of a mutation in *CTNNB1* [97], *ZNRF3* and *RNF43*, *RSPOP2* [85]. Many mutations described above are target of specific drug treatment and used as biomarkes to design personalized treatments for patient suffering from advanced CRPC [83, 85, 98].

2.3 Intrinsically disordered proteins

Early observations on the relationship between enzymes structure and function are from 1961 when random coil behavior and loss of enzymatic activity was observed in bovine pancreatic ribonuclease [99]. From this and subsequent observations, it was postulated that protein structure determines function and its alterations hinder enzymatic activity. Although this paradigm holds for many proteins, since the 90s a novel class of protein lacking fixed three-dimensional structure has been characterized and, in 1999, P.E. Wrigth and H. J. Dyson proposed to re-assess the traditional structure-function paradigm in light of these new observations [100]. These proteins, in physiological conditions, do not have a globular structure and behave like random-coils [101] or as an ensemble of inter-converting conformers [102, 103, 104, 105, 106] (Figure 2.3). This class of proteins has been called intrinsically disordered proteins (IDPs) and their biological function is tightly linked to their biophysical properties [100, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116]. Activation of calcineurin upon interaction between an intrinsically disordered region (IDR) and the Ca^{2+}–calmodulin complex was one of the first examples of this phenomenon [107, 108]. Early observations on histone N-terminal tail suggested that acetylation reduces rigidity [107]. These early results were generated with *in vitro* systems, thus whether this phenomenon happens *in vivo* was soon addressed. The proto-oncogene c-Fos and the cell-cycle inhibitor $p27^{Kip1}$ both have IDR in the domains used for molecular interactions and have been shown to maintain their flexibility also in crowded environment, like cell nucleus [109]. In viruses, IDRs are involved in assembly of macromolecular complexes, e.g. the TMV particle nucleation process is initiated and stabilized by coat proteins with negatively, highly flexible IDRs facing the inner cavity of the nascent rod-shaped viral particle and interacting with the single stranded viral RNA; other examples are the assembly of icosaedral viruses and, in bacteria, assembly of the flagellum [110]. In humans, the presynaptic protein α–synuclein, associated with Parkinson disease and insurgence of other neurological disorders, lacks of a rigid globular structure and can fold in multiple conformations [111]. Also the N-terminal domain of many nuclear hormone receptors display high flexibility and have been shown to change conformation upon interaction with small molecules (e.g. hormones) [106]. Transcriptional co-activators CBP and p300 acetylate histones and stabilize molecular interactions between TFs and the transcriptional machinery

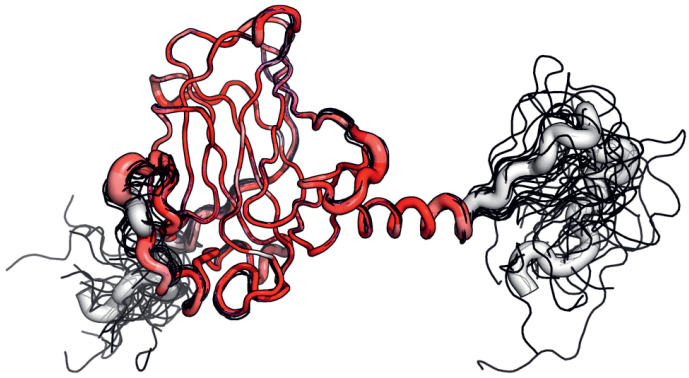


Figure 2.3 Cartoon representation of TP63 DNA-binding domain. Backbone of the protein is represented as a tube. For this drawing, structure from 61 homologous PDB entries were superposed and the tube size is proportional to the root mean square deviation (RMSD) per residue between C-alpha pairs. The white to red color ramping is used to visualize sequence conservation. Conformers from liquid NMR experiment are displayed as black traces. Rendered from PDB structure 2RMN [125] with ENDscript 2.0 [126].

while displaying more than 50% of residues in IDR [105]. Many more examples could be listed, and, among the others TP53, chaperon proteins and BRCA1 have been characterized to have at least one IDR [112].

As show above, intrinsic protein disorder is present in all kingdoms of life but is enriched in eukaryotic organisms and displays a positive correlation with organism complexity. About a third of eurkaryotic proteins display at least one intrinsically disordered region (IDR) [117, 118, 119, 120]. It has been shown that intrinsic protein disorder arises at a later stage of the evolutionary process [121] and could be linked to more complex molecular functions required by eukaryotic cells for their functioning. Three quarters of proteins mutated in human cancers are estimated to have at least an intrinsically disordered region (IDR) [122, 123, 124].

Intrinsically disordered proteins represent a major component of the dark proteome [127]. This term is used to describe the subset of protein universe whose three-dimensional structure has never been observed. It has been estimated that more than half of the proteins in higher eukaryotic proteomes is constituted by at least one unobserved IDR [128].

Since early studies, sequence composition of IDP appeared to be biased [103, 120, 129, 130, 131]. Some residues have been associated with intrinsic protein disorder and thus defined disorder-promoting residues (Pro, Glu, Lys, Ser and Gln). They are

characterized by net charges and reduced hydrophobicity. While showing enrichment for disorder-promoting residues, IDP show depletion of structure-promoting residues (Trp, Tyr, Phe, Cys, Ile, Leu and Asn) [107, 132, 133].

2.3.1 Biological functions

Absence of a defined three-dimensional structure makes these proteins unsuitable for enzymatic functions but entails them to function as regulators of many biological processes including transcription and cell cycle [105, 112, 134, 135].

The functional classification of IDP has been a major topic of discussion since early reports. Five broad functional categories were first proposed [101] and this classification has been later refined with the addition of newly discovered classes [136]. Currently, the six major ones are: entropic chains, display sites, chaperons, effectors, assemblers and scavengers:

- Entropic chains were the first observed type of intrinsically disordered regions (IDRs), they can be described as protein regions that carry out functions which directly benefit from conformational disorder [136]. Linkers between globular domains, loops and spacers are typical examples of IDRs with entropic chain function.
- Display sites are IDRs targetd by PTM. Their flexibility facilitates the modification deposition, inducing an energy loss that allows the interaction with other proteins. These regions are well studied because of their intimate involvement in cellular signaling [122, 135, 137, 138, 139].
- Chaperons are proteins that help other proteins or RNA to fold properly. Enhanced flexibility helps chaperons to adapt to many binding partners and enable fast intermolecular interaction.
- Effector proteins interact with other proteins and modify their behavior. IDP with this function, sometimes, can alter the activity of other parts of the same protein.
- Assemblers take part in the creation of higher order molecular complexes. Proteins with this function have multiple IDR that concurrently bind different partners helping to bring together subunits of large complexes.

- Scavenger proteins bind and neutralize small ligands. Their role is to regulate ligand availability for other molecules.

Display sites, chaperons, effectors, scavengers and assemblers share the fundamental function of molecular recognition. Prior or upon interaction they may undergo structural modifications that induce an entropic loss resulting in a disorder-to-order transition [140, 141, 142]. The unbound protein in its disorder state and the reduction in entropy that drives the folding process are the key factors that regulate these interactions [101]. In other words, these regions may fold upon binding and the likelihood of this process has been correlated with the secondary structure elements present (or predicted) in the protein sequence [121]. Using intrinsic disorder for molecular recognition has some additional benefits: first, these proteins show rapid association/dissociation kinetics, which allow for rapid response to external stimuli. Second, since their backbone is extremely flexible they can adapt to multiple interaction partners and thus be involved in many regulative patways. Intrinsically disordered proteins interaction promiscuity is allowed by the large number of conformations the unbound state of the protein can take and the ability to fold in different ways upon upstream stimuli. Because of this, these proteins tend to occupy a central positions in biological networks. IDPs, often act as stimuli integration hubs. Adaptation to multiple cellular environments [140, 143] and interaction with proteins from different signaling pathways makes them able to integrate stimuli into coherent responses. These proteins represent the conserved core of protein signaling networks, are responsible for signal integration and altogether constitute the ability of a cell to react and adapt to multiple stimuli [144]. Moreover, in some instances, the disorder state is maintained upon molecular interaction [101].

Because of their critical role, the intracellular concentration of IDPs is lower than globular proteins. In their unbound disordered state, they are susceptible to proteolytic cleavage. Moreover, IDP transcripts tend to have more predicted miRNA binding sites and ubiquitination sites as well as higher decay rates [145]. Moreover, dosage sensitivity has been associated with intrinsic protein disorder: many dosage-sensitive genes have been shown to code for proteins with extensive IDRs [146]. From these observations emerges that IDP have short half-lives and are present at low concentration in the cell. In some cases, however, IDPs get stabilized by interactions with other molecules, leading to avoidance of proteasomal degration, thus allowing the creation of multimeric functional complexes [147].

To summarize, IDP are a class of proteins which, in solution, lack of defined three dimensional structure. This feature makes them well suited for signaling and molecular recognition functions. Many examples folding-upon-binding IDP exist, but this phenomenon is not observed in all cases. These proteins tend to occupy central positions in signaling and protein-protein networks and because of this their intracellular concentration is carefully controlled.

2.3.2 Role of intrinsic protein disorder in gene expression regulation

Transcription factors structural features are involved in binding specificity, regulation and sensing. For instance, the N-terminal domain of TP53 is annotated as IDR. It binds TP53 DNA binding domain blocking unspecific interactions with the DNA. This self-inhibition boosts the specificity of protein-DNA interactions [148]. Other TFs display IDRs outside the DNA binding domain which have been implied in directing binding site recognition, thus regulating site-specific selection [2].

transcription factors are key regulators of eukaryotic gene expression. Structural organization of these proteins is substantially conserved: a globular DNA binding domain and an activation domain characterize the structure of a vast majority of these proteins. Activation domains are involved in interactions with other TFs or small ligands and are characterized by low-complexity, flexible, IDR. Mutations in these domain abolish transcription and may give rise to pathological phenotypes [149]. The interaction among activation domains not only activate the TF but stabilizes DNA binding, interactions with cofactors, recruitment of the polymerase complex and activation of the transcriptional process [150].

Enhancer-bound TFs recruit the Mediator complex and other cofactors to activate gene expression at promoters. Super-enhancer are genomic loci with higher density of enhancer elements and TF binding sites. It has been reported that the binding of a TF on these loci, induces recruitment of the Mediator complex and BRD4. Formation of this complex is driven by weak interactions among IDR from the enhancer-bound TFs, MED1 Mediator subunit and BRD4. As a result, phase-separated droplets form at these loci [19] (Figure 2.2B). Within these temporary nuclear sub-compartments RNA polymerase subunits can diffuse and the transcription machinery assembled [19]. These findings were confirmed with the OCT4, GCN4 and ER TFs [151]. Enhancer propensity to form these condensates is not only encoded in IDRs sequences

but also in the number of binding sites composing the enhancer, in the strength of protein-DNA interaction and in the TF (and cofactors) intracellular concentration [152]. DNA binding is required to stabilize droplets and its formation stabilizes weak IDR-IDR interaction as reported by thermodynamic analysis [152]. This mechanism suggests that the cooperative role of all molecular species involved is important to achieve correct biochemical composition and precise genomic localization of trascriptional condensates.

2.3.3 Computational methods for intrinsic protein disorder detection and prediction

Given the distinctive features of IDPs, the challenges in obtaining three-dimensional models of their structures and their existence as a structural continuum of conformers, a plethora of prediction methods have been developed. Each tool uses a different approach to predict this property, and they can be grouped in three main categories: biophysical properties-based methods, machine learning-based methods and meta-predictors (Table 2.1). In the MobiDB [153, 154] database fifteen different tools are used to predict intrinsic protein disorder and secondary structure populations for the entire protein universe:

- Mobi 2 [155] annotates protein sequence with mobility and ID information from missing electron densities, high B-factor (X-ray and electron microscopy) and inter-model mobility from NMR ensembles. Identifies also linear intercting peptides (LIPs).
- δ2D [156] uses backbone chemical shifts from NMR-resolved structures to predict populations of secondary structures and define protein states (fully structured, partially folded, disordered).
- Random Coil Index (RCI) [157, 158, 159] quantifies the propensity of a polypeptide to assume a random coil conformation using NMR chemical shifts. The method relies on an empirically determined equation.
- IUPred [160, 161, 162] uses a manually curated table of pairwise energy values to compute probability of a residue to lie in an IDR [176].
- Anchor [163, 164] is a specialized predictor to identify disordered segments able to undergo disorder-to-order transition. The prediction process relies on

Table 2.1 Overview of intrinsic protein disorder prediction methods.

Name	Reference	Predicted feature	Prediction model
Mobi 2	155	IDRs, LIPs	Biophysical properties
δ2D	156	IDRs, secondary structure populations	Biophysical properties
RCI	157, 158, 159	IDRs	Biophysical properties
IUPred	160, 161, 162	IDRs	Biophysical properties
Anchor	163, 164	IDRs, disorder-to-order likelihood	Biophysical properties
ESpritz	165	IDRs	Machine learning
FELLS	166	Secondary structure populations	Machine learning
RING 2.0	167	Intra- and inter-chain interactions	Biophysical properties
DisEMBL	168	IDRs	Biophysical properties
GlobPlot	169	IDRs, secondary structure elements	Biophysical properties
RONN	170	IDRs	Machine learning
VSL2b	171, 172	IDRs	Machine learning, meta-predictor
SEG	173	Low-complexity	Biophysical properties
Pfilt	174	Low-complexity	Biophysical properties
Dynamine	175	Backbone flexibility	Machine learning

previous identification of IDR with IUPred [160, 161, 162]. The classification is based on two more criteria: first it calculates the number of inter-molecular contacts a residue can make with neighboring residues to ensure it cannot fold, then it calculates the number of favorable intra-molecular contacts with the interaction partner to ensure there is an energy gain in the interaction and thus the ability to fold.

- ESpritz [165] refer to an ensemble of four predictors using a bidirectional recurrent neural network (BRNN) [177] to predict intrinsic protein disorder

from sequence alone. Different tools are trained on different data sets to predict intrinsic protein disorder derived from specific experimental techniques: X-ray from PDB, DisProt, NMR and MxD [178]. A final consensus prediction is computed by averaging predictions from the separate tools.

- FELLS [166] aggregates structural predictions and sequence propensities from different sources: Espritz–NMR [165] and a method derived from the Espritz neural network architecture called FESS. This is an alignment-free method based on bidirectional recurrent neural network (BRNN) [177].
- RING 2.0 [167] identifies residue-residue interactions via analysis of RINs derived from PDB structures. It is able to identify inter and intra chain covalent and non-covalent bonds, π–π stacks and π–cation interactions.
- DisEMBL [168] defines ID from a two–states model of protein structures: each residue can either be ordered or disordered. The state assignment is performed based on three criteria: DSSP [179] secondary structure prediction, high B-factor and X-ray missing electron density.
- GlobPlot [169] classifies protein residues in two states: random-coil and secondary structure. It uses a scale computed from Russel/Linding propensity scale [180] and DSSP [179] secondary structure prediction from a set of representative proteins selected from SCOP [181, 182]. For an input protein sequence, for each residue, a cumulative score is computed using the propensity scale and the classification is performed by peak detection over the computed signal. Peak detection assigns the "random-coil" class if the signal function derivative is positive and "secondary structure" otherwise.
- RONN [170] uses a bio-basis function neural network (BBFNN) to compute the probability of a fixed-size stretch of amino acid to be disordered. The prediction is based on the computation of a distance value between the input query and a set of known prototype sequences. The classification is then made according to the closest (most similar) prototype sequence.
- VSL2b [171, 172]: is a meta-predictor combining the outputs of two SVM-based predictors (VSL2-L and VSL2-S) trained independently to detect long and short stretches of disordered residues. The input features are based on statistical, physico-chemical and evolutionary properties of the protein sequence. The meta-predictor takes as input the disorder probabilities computed from the

two components and outputs the class probability of observing disordered or ordered state. This meta-predictor is trained independently of the other two components.

Moreover, in MobiDB, low complexity regions and backbone flexbility are predicted using dedicated tools:

- SEG [173] is one of the first algorithms developed to predict low-complexity regions. It uses sequence composition of a fixed length window to predict its complexity score.
- Pfilt [174] is an algorithm designed to mask out regions of low complexity, coiled-coil regions and regions with extremely biased amino acid compositions. It was developed to control for error rate in PSI-BLAST alignments and improve on the SEG family of algorithms.
- Dynamine [175] computes proteins backbone dynamics using a linear regression model and a sliding window approach to achieve residue-level flexibility prediction.

2.4 High throughput sequencing methods

DNA sequencing is a technique to determine the sequence of nucleotides forming a DNA molecule. The first proposed experimental procedure was the Sanger method developed in 1975 by Dr. Frederick Sanger [183]. This leverages on DNA polymerase and radiolabelled nucleotides to manually reconstruct the sequence of a given DNA molecule. The main limitation of the Sanger approach was its throughput, the manual intervention needed to reconstruct the input sequence was a limiting factor to the quantity of analyzed genetic material.

Thanks to the momentum generated by the Human Genome Project (HGP) [184, 185, 186], in the second half of the 90s a series of new technologies emerged improving the Sanger approach. Main improvement was the shotgun method. In summary, this technique requires random fragmentation of the original genetic information, amplification and division into smaller overlapping segments, sequencing and sequence reconstruction by assembling the read data into the original segments. To generate sequence data for the HGP, the pyrosequencing method was introduced [187]. The

novelty of this method was the "sequencing by synthesis" approach that allowed to read a DNA strand concomitantly to its synthesis. In brief, upon insertion of a known new base, a detectable light signal is emitted. An optical sensor detects such signal allowing the reconstruction of the input sequence.

After the completion of the HGP, more efficient and cost-effective sequencing methods have been developed. They leverage on the knowledge generated during the project to further improve the sequencing throughput. Collectively these methods are called next-generation sequencing (NGS) methods. Capillary elettrophoresis was developed for the HGP to parallelize the Sanger and pyrosequencing methods [188]. Furthermore, the Illumina dye method improves on parallelization using dense chips of anchored oligonucleotides and an improved chemistry to synthesize template DNA molecules *in-situ* [189]. The sequencing reaction is carried out using a sequencing by synthesis approach. Pacific Biosystems developed a real-time single molecule sequencing technology. This system works using immobilized DNA polymerase on top of a detector sensing labeled nucleotides as they get inserted into a nascent DNA strand [190]. Other methods were proposed and involved different approaches, for instance Applied Biosystems developed the SOLiD® technology implementing a sequencing by ligation technique. This technology relies on a ligation reaction between a known sequence fragment and a labelled oligonucleotide reporting two known bases [191]. These technologies paved the way to large genomic, transcriptomic and metagenomic studies while changing the way classical subjects like molecular biology, genetics and virology are studied.

Third generation sequencing technologies have been deployed in recent years. These machines improve on read lengths, portability and spectrum of applications. Pacific Biosystem, improving on their previous real-time single molecule system, leads the competition with the Oxford Nanopore Technology which is able to produce reads as long as few kilobases. The idea of Nanopore sequencing is to use a biological pore of known diameter to feed a single nucleic acid molecule to a polymerase molecule, then read the nascent DNA or RNA molecule in a sequencing by synthesis fashion [192, 193].

2.4.1 Transcriptome sequencing

RNA-sequencing (RNA-seq) was introduced in 2008 by independent authors studying different eukaryotic systems (mouse, yeast and Arabidopsis) [194, 195, 196] during the rapid expansion of NGS technologies. The introduction of this technology marked a big leap into gene expression quantification by simplifying the experimental procedure and reducing costs. Although the most popular application of RNA-seq method is the quantitative analysis of transcriptome, this term can refer to any kind of sequencing involving RNA.

NGS platforms are the most suited for RNA-seq studies as the amplification step they require yields products in a concentration proportional to the original RNA molecule concentration. Data analysis require assembly of reads against a reference transcriptome before quantification.

According to the type of RNA under analysis, RNA-seq libraries are prepared in different ways. RNA-seq library preparation requires an enrichment step to isolate such molecules, smallRNA-seq a size selection step to enrich for short mature and precursor miRNA. In either case a retrotranscription step is required to produce cDNA before actual sequencing.

Traditional computational analysis involved artifacts and erratic reads removal. Reads shorter than a fixed length can be discarded at this stage. Sequencing adapters trimming is another integral part of the pre-processing procedure although questioned in recent times [197]. After preprocessing, reads need to be aligned to a reference genome. This step provides evidence of transcribed genomic regions. To quantify gene expression, alignment to transcriptome and downstream transcripts quantification is required. Optionally, quantified transcripts can be collapsed at gene level and used for differential expression analysis. For smallRNA-seq analysis, miRNA quantificaion is run using an encyclopedia of known miRNA sequences. Some tools are able to estimate miRNA abundance taking into account also precursor sequences [198].

Usually, RNA-seq experiments are run to compare gene expression across different conditions. Genes that change their expression level significantly are called differentially expressed genes (DEG). Over the years, multiple methods have been developed to detect such genes and a wide variety approaches have been published [199, 200, 201, 202]. Alignment-free methods have been also proposed for differential gene expression analysis [203, 204].

Irrespective of the quantification method, a filter for lowly expressed genes is good practice for RNA-seq data analysis. The empirical distribution of mean (or median) expression values can be computed from observed data, quantile values can be calculated and used as thresholds to remove lowly expressed genes. Moreover, sensitivity and specificity of differential expression detection can be tuned with few parameters. First, $\log_2$-fold change is computed to describe gene expression change across the two conditions:

$$\mathrm{FC}_i = \log_2\left(\frac{\overline{G_{1i}}}{\overline{G_{2i}}}\right) \tag{2.1}$$

Where $\overline{G_{1i}}$ and $\overline{G_{2i}}$ represent the average expression of gene G_i in the two conditions. Moreover, the null hypothesis of no difference between condition is usually tested. Commonly used tests are t-test and Mann Whitney U test, in either their paired or unparied variants depending on the data. A significant p-value leads to reject the null hypothesis implying that the observed difference is not due to randomness or chance. Additionally, absolute median difference between estimated gene counts can be computed:

$$\mathrm{diff}_i = \overline{G_{1i}} - \overline{G_{2i}} \tag{2.2}$$

For these values, an empirical distribution can be estimated and a filtering procedure applied to remove genes showing low difference across conditions. This extra step increases the signal-to-noise ratio by removing genes with small average expression values but high $\log_2$-fold change ratio.

2.4.2 Sequencing methods for chromatin structure and epigenetics study

Study of epigenetics focuses on histone modifications, DNA methylation, the interaction between TFs and genomic DNA as well as all the proteins involved in these mechanisms. In the pre-genomic era main methodology to study these interactions was chromatin immunopurification (ChIP) which relied on antibodies-driven enrichment of DNA-bound factors and subsequent chromatograpic purification. With the advent of sequencing methodologies, the throughput of these assays was improved by coupling them with sequencing (ChIP-seq) [205, 206]. From analysis of ChIP-seq

data, TF binding sites and histone modifications can be mapped on target genomes. This technique was used in large projects to define genome-wide modifications maps, e.g. ENCODE [207] and ROADMAP epigenome [208]. Main limitation of ChIP-seq is that it can probe a single TF or histone modification at the time. Because of this, recent techniques have been proposed to simultaneously quantify binding of multiple TFs, e.g. CUT&RUN [209], ChIP-exo [210] and SELEX-seq [211].

Chromosome conformation capture experiments provide insights into the pattern of genomic DNA binary contact. According to the number of probed interactions 3C, 4C and 5C experiments can be distinguished: 3C techniques test one-vs-one interactions allowing the study of specific interaction between an enhancer and a promoter [212]. To identify all enhancer elements able to bind a specific promoter, 4C, also known as capture-on-chip techniques can be used [213, 214]. To study all possible interaction within a given genomic region, 5C techniques have been developed [215]. Similarly to ChIP-seq the main limitation of these techniques is throughput, none of these is suitable for genome wide studies. Hi-C is a technique that overcome this limitation and uses paired-sequencing to generate genome wide contact maps allowing the identification of chromosome territories, TADs and loops [216].

The assay for transposase-accessible chromatin coupled with sequencing (ATAC-seq) approach has been proposed in 2013 [217]. The assay involves utilization of a hyperactive viral Tn5 transposase to detect regions of accessible chromatin before performing sequencing. Tn5 protein binds a linear DNA molecule and catalyzes its insertion into host genome. In physiological condition the transposase is loaded with a transposon sequence and catalizes its insertion with a cut-paste mechanism. For this application the transposon sequence is replaced by adapters so that, while the transposase catalyzes their deposition in nucleosome-free genomic regions, it introduces a double strand break. This process generates DNA fragments suitable for sequencing. Average length of these fragments is proportional to the length of DNA stretch wrapped around nucleosomes. By analyzing the length distribution and the aligned reads pattern, a detailed map of Tn5 accessible chromatin regions and nucleosome position can be computed. This experimental approach can be used to characterize regulative regions in genes promoters, enhancers and silencers, if annotations are available. Moreover, from the analysis of peaks shape, TFs occupancy can also be inferred: if a TF is bound to DNA, a footprint can be detected in the

signal generated by the aligned reads. ATAC-seq is a powerful approach to study the regulatory epigenome and its role in regulating gene expression. For instance, this technique was used to characterize the cis-regulatory elements of the whole TCGA cohort and the resulting data integrated with transcriptomic and other data type to identify enhancers active in primary human cancers [218].

3 AIMS OF THE STUDY

The general aim of my PhD project was to study biomolecules from their structural perspective: first how the DNA structure contributes to the regulation of gene expression in prostate cancer, and, second contribute to the knowledge on intrinsic protein disorder by providing the research community with tools to describe and study it. In the light of these broader aims, the aims of my publication can be summarized as follows:

1. Characterize the relationship between chromatin accessibility and gene expression during prostate cancer progression by defining a group of relevant regulative loci
2. Contribute to the study of intrinsic protein disorder by deploying web–based resources to store and visualize literature–derived manually-curated and large–scale automatically–generated ID structural annotations;
3. Identify regulative programs leading prostate cancer progression by detecting transcription factor binding sites within putative regulative loci, analyze co-binding patterns and characterize their structure with a computational approach and developed resources;

4 MATERIALS AND METHODS

This chapter will be split in two main sections: the first will describe samples and methods used to run the analysis presented in Publication I, while the second will describe the overall organization of applications presented in Publication II and Publication III.

4.1 Tampere PC cohort (Publication I)

The study was performed on samples collected at the Tampere University hospital. This material is part of the Tampere PC cohort which is made of freshly frozen uncultured tissue biopsies from BPH, PC and CRPC patients [66]. The cohort mimics PCa progression: BPH represents the normal control, PC the first cancer stage and CRPC the advanced, treatment resistant stage.

Samples were collected using two different surgical procedures: radical prostatectomy (RP) and trans-uretral radical prostatectomy (TURP). The former involves complete removal of the prostate gland, while the latter is a minimally invasive procedure that involves insertion of a needle in the urethra to reach the pelvic cavity and sample the prostate epithelium. TURP is known to induce ischemic damage to the bioptic material which results alterations in gene expression [219]. Given the mixture of sample collection methods, a substantial known batch effect is present and will have to be handled in downstream analyses.

Samples from the Tampere PC cohort have been extensively characterized over the years: shallow whole genome sequencing, RNA-seq, methylation and proteomics data have been generated and analyzed [220, 221, 222, 223]. For the project presented here, chromatin accessibility data (ATAC-seq) have been generated.

The composition of the cohort changed over time. In the first publication, in 1995 [66], 6 BPH, 31 PC and 9 CRPC were used for the study. In 2013, for the original RNA-seq study [221], 12 BPH, 28 PC, and 13 CRPC samples were utilized.For the

proteogenomic paper published in 2018, [222], 10 BPH, 17 PC, and 11 CRPC samples were used.

4.1.1 RNA-seq

Gene expression quantification. Total RNA was extracted from 45 samples of the Tampere PC cohort and used to generate sequencing libraries for RNA-seq [221]. Of these, 37 had ATAC-seq data available: 10 BPH, 16 PC and 11 CRPC. Sequencing reads have been aligned to human genome version 38 (GRCh38) with STAR version 2.5.3a [224] and gene counts were computed for 58,243 genes from Ensembl version 90 (August 2017) using `GeneCount` run mode from STAR. Unstranded, sample–specific counts were collected in a 58,243×37 data matrix G for further processing.

Preprocessing. For each gene, a vector g_i of total gene expression values was calculated by summing over columns of matrix G:

$$g_i = \sum_j G_{ij} \tag{4.1}$$

where i represents genes and j represents samples. The empirical distribution of total gene expression was computed and the value corresponding to its lower quartile q_{25} computed. Genes with total gene expression lower than q_{25} were discarded from further analysis. This procedure yielded 18,537 genes above threshold.

Removed genes show a consistently low expression across all samples. Because of this, they have a very small information content for the biological process under investigation (PCa progression). Thus, removing these genes enhances the signal-to-noise ratio and improves the downstream differential expression analysis by removing a source of unwanted noise. Moreover, genes with low read count are likely to suffer from overdispersion, a phenomenon by which the observed read count variance is substantially larger than the expected variance under certain statistical model. To account for this phenomenon, DESeq2 [200] models gene expression using a negative binomial distribution which depends on a dispersion parameter. The DESeq2 model implements a Bayesian method to estimate such parameter from the input data. Thus, by providing data with higher signal-to-noise ratio, dispersion estimation procedure yields more accurate estimates improving the overall quality of the results.

Differential expression analysis The prefiltering procedure yielded a 18,537×37 matrix of filtered gene expression values. It was used as input for DE analysis with DESeq2 [200]. A design matrix modeling both the sample type (BPH, PC, CRPC) and the collection method (TURP, RP) was generated to represent experimental variables.

After model fitting, for each experimental variable, model coefficients were retrieved. Next, to address the problem of TURP–derived batch effect, coefficients computed for the "collection method" variable were extracted and used as correction factors:

$$N_{ij} = \log_2\left(\frac{G_{ij}}{\sum_i G_{ij}} - \beta_{TURP} X^T\right) \tag{4.2}$$

where N_ij is the normalized value of gene i in sample j, $\sum_i G_{ij}$ the library size of the j^{th} sample, $\beta_{TURP} X^T$ is the matrix of gene-specific β coefficients estimated by the model multiplied by a binary vector X of length j whose values are 1 for samples collected with TURP and zero otherwise.

DESeq2 automatically computes $\log_2$ fold–change and p-value for any given comparison. On top of these values, a vector of absolute median difference was computed:

$$\text{diff}_i = \overline{N}_{i\,g_1} - \overline{N}_{i\,g_2} \tag{4.3}$$

where diff_i is the median difference for gene i, $\overline{N}_{i\,g_1}$ and $\overline{N}_{i\,g_2}$ are the median values in sample groups g_1 and g_2 from the N matrix. This metric introduces an extra requirement to achieve the status of DEG.

The median $\log_2$ fold–change value describes how different normalized median gene expression values are. If calculated between small numbers, it results in high values and thus, false positive DE calls. In other words, genes with barely detectable expression but different enough values in the comparison get called as DEG. Using the absolute median difference introduces and extra constrain helping in removing these edge cases and thus improving the detection power. This procedure yields DE calls with strong, detectable and clearly different signal between conditions.

To label any gene as DE, the following three criteria were used: first, the $\log_2$ fold-change was required to be greater than 1, FDR-corrected p-value to be lower than 0.05 and absolute median difference to be greater than 180. Two sets of DEG were computed: BPH vs. PC and PC vs. CRPC. In the first 933 genes were detected, 508

of which were upregulated and 425 downregulated. In the second, 533 were detected, 194 were upregulated and 339 were downregulated.

4.1.2 SmallRNA-seq

miRNA expression quantification. To run smallRNA-seq experiment, a size selection step is performed before retrotranscribing RNA to DNA. This allows to select RNA molecules of low molecular weight corresponding to miRNA and small non-coding RNA. For this analysis, only miRNAs were quantified. To quantify miRNA expression, sequencing reads were mapped to human mirBase database version 22 [225] allowing for single base deletion at 3' or insertion at both ends. Expression values for miRNA were computed by summing reads mapping either to mature sequences or precursors. The resulting matrix was normalized with median-of-ratios normalization. This procedure yielded quantification for 1,471 miRNA. The matrix of normalized counts was then used to detect DE miRNA.

Differential expression analysis. Differential expression was detected for the two comparisons described previously. Similarly to mRNA-seq, $\log_2$ fold–change was required to be greater than 1, FDR-corrected t–test p–value to be lower than 0.05 in order to call a miRNA DE. No absolute median difference was used because of the reduced number of quantified molecules. For BPH vs PC and PC vs CRPC, 26 and 51 DE miRNA were called respectively.

4.1.3 ATAC-seq

From ATAC-seq experiment, peaks and differentially accessible regions (DARs) were detected (Publication I). Normalized ATAC-seq signals were analyzed together with RNA-seq and smallRNA-seq data to identify regulatory regions relevant in PCa progression. First, target genes for candidate regulatory regions were predicted. Next, leveraging on publicly available ChIP-seq data, TF binding sites content of said regions was analyzed to identify groups of TF involved in regulation of gene expression over PCa progression, finally the identified TFs were used to build a regulative network. Unsupervised clustering of this network yielded groups of TFs co-regulating large numbers of genes (Figure 4.1).

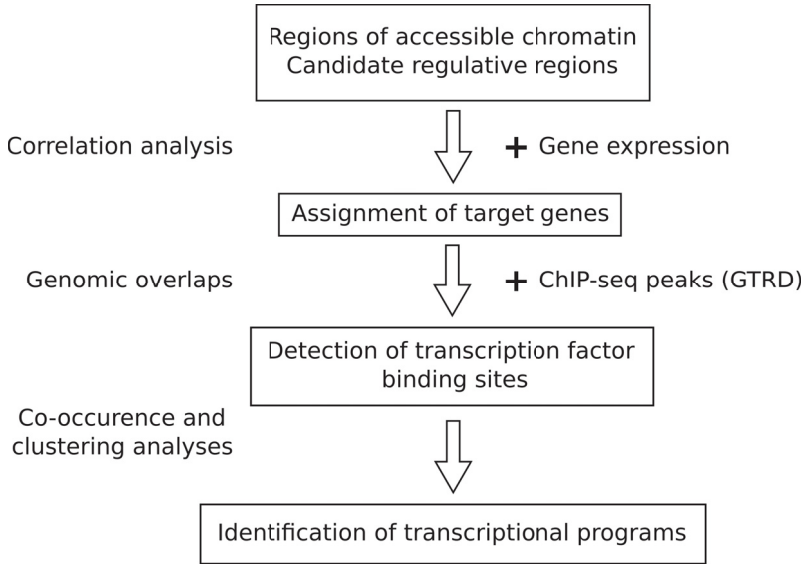


Figure 4.1 Worflow of the analysis performed to identify transcriptional programs from ATAC-seq, gene expression and ChIP-seq data.

Integrated analysis with RNA-seq. To assign candidate regulatory regions to genes, Pearson and Spearman correlation coefficients between ATAC-seq and gene expression signals were computed. A regulatory region was assigned to a gene if the absolute value of any of the two correlation coefficients was greater than or equal to 0.5.

To compute correlation coefficients, samples shared among the two experiments were used: if the gene was a protein coding gene, samples shared between RNA-seq and ATAC-seq data set were used, otherwise, if the gene was a coding for a miRNA, samples shared between ATAC-seq and smallRNA-seq data were used. Because of this limitation, the sample size available for correlation analysis with miRNA was lower and thus false positive ratio slightly higher.

Values of Pearson correlation coefficient tend to drop when the amount of variation between the two signals is not the same, e.g. the ATAC-seq signal varies a lot between samples, while the RNA-seq does not or vice versa. In these cases, if the data trend is consistent (i.e. the two signals vary consistently), the Spearman correlation coefficient reports a substantial correlation because it quantifies the covariation between data ranks instead actual values. Combining the two coefficients allows to include these corner cases at the expense of a higher false positive rate.

To control for increased false positive rate, a high threshold value was used. To quantify it, null distributions were computed randomizing the signal values and computing new correlation coefficients in all the genomic contexts described below. Through this procedure, the threshold value was set to the above mentioned value (0.5).

For each gene, candidate regions were searched in different genomic contexts (Figure 4.2):

- **Gene's TSS**. ATAC-seq signal was quantified in a 1 kbp area centered on gene TSS and normalized according to the DARs pipeline (Publication I). Correlation values were computed between such accessibility value and expression level of the annotated gene (Figure 4.2A).
- **Gene's promoter area.** For each gene, an asymmetric promoter region of 1.1 kbp spanning 1 kbp upstream to 100 bp downstream of the TSS was defined. Normalized intensity values of ATAC-seq peaks and DARs overlapping this region were used to compute correlation with gene expression level (Figure 4.2B).
- **Closest gene to each ATAC-seq feature.** Genomic distance between each peak or DAR and their closest gene was computed with HOMER [226]. Normalized intensity value of ATAC-seq features and closest gene was computed (Figure 4.2C).
- **Genes and ATAC-seq features within the same TAD.** TADs boundaries from PCa cell lines were retrieved from publicly available data [227, 228]. Genes and ATAC-seq features localizing to the same TAD were detected and correlation between all couples computed (Figure 4.2D).

Integrated analysis with ChIP-seq. To characterize transcriptional programs driving PCa progression, TF binding sites were searched in DARs with accessibility signal showing strong correlation with gene expression. Genomic locations of TF binding sites detected with ChIP-seq were retrieved from GTRD [229]. This database collects uniformly processed peaks from ChIP-seq experiments ran in cellular models. In brief, to generate this data, ChIP-seq reads were collected from different source and aligned to the same genome version. Peaks were detected with four different methods and clustered according to experimental conditions (e.g. cell line or treatment). For

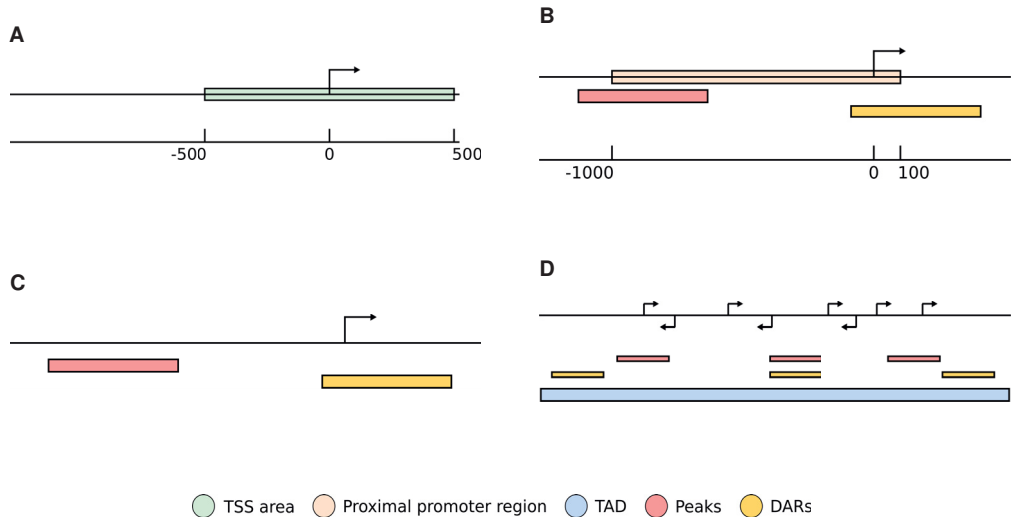


Figure 4.2 Schematic representation of genomic contexts used to assign genes their candidate regulatory regions. **A.** Correlation between gene expression and chromatin accessibility signal quantified in the TSS region. **B.** Correlation between gene expression and peaks or DARs overalapping the proximal promoter region. **C.** Correlation between ATAC-seq peaks or DARs and their closest annotated gene. **D.** Correlation between all genes and peaks or DARs within the same TAD.

clustering, peaks got reduced to their centers, i.e. the center base of the binding site and, finally, clusters for the same TF were joined into metaclusters. Cluster located 50 bp away from each other were grouped to generate the set of binding sites used in this analysis. If a binding site from GTRD overlapped any DAR identified during the correlation analysis, the TF was reported to be present in that feature. The overlap procedure was implemented in R 3.5.2 using the `findOverlaps` function from the `genomicAlignments` package (Bioconductor 3.5.2).

Building of a regulative network. From previous analyses a set of genes and a set of accessible genomic features were identified. In these features TF binding sites have been detected and a putative regulative function inferred. To discover groups of TF co-regulating gene expression, a regulative network was built. For each TF, the list of genes correlating with accessibility of features reporting a binding site was computed, resulting in a list of putative regulated genes. For each couple of TFs the number of shared genes was computed resulting in a contingency matrix. Couples of TFs sharing less than 300 genes were discarded. From this filtered matrix, an undirected weighted graph was built (Figure 5.2). Transcription factors were represented as nodes

and the number of shared putatively co-regulated genes encoded in the edge weight. In order to visualize it, edges were pruned and only the upper half of the edges weight distribution retained. Edge color and width was set to reflect the weight value and interesting genes were highlighted by increasing the node size (e.g. AR).

4.2 Web resources for intrinsic protein disorder annotation (Publication II, Publication III)

The general structure of the two software applications developed for intrinsic protein disorder annotation is very similar (Figure 4.3), both of them are composed of three software layers: a data layer powered by non-relational DBMS storing protein annotations, a back-end layer querying the database and serving query results via REST endpoints and a front-end layer making use of the back-end to display data via a standard web browser. This separation of concerns improves modularity and code re-usability, allowing to decouple the development of all components.

4.2.1 Databases

IDP annotations for both DisProt and MobiDB are stored in non-relational DBMS. The engine of choice for both the applications is MongoDB 3.8. Data are stored as JSON objects (also called documents) which are constituted of key-value pairs. Sets of related documents are grouped in collections. In MongoDB, documents are unrelated entities, thus, no joining operation can be performed across collections or within a single collection of documents.

DisProt DisProt annotations are stored in an IDR-centric fashion: each IDR is represented as a separate document within a collection. Each document has a unique internal identifier and carries information about IDR localization on the protein sequence. Moreover, an identifier for the protein of origin, experimental detection method, reference to the experiment reporting paper, identifier for the person that curated the annotation and timestamp are stored. Each IDR belongs to an IDP and each IDP can have multiple IDR of variable length. To handle such redundancy, a separate collection of entries representing proteins was designed. These, store

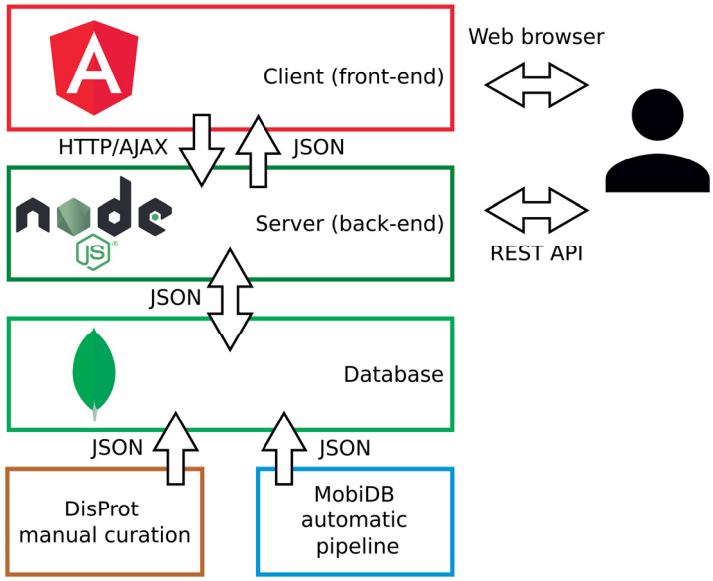


Figure 4.3 Schematic representation of software stack used for IDP databases. Intrinsic protein disorder annotations are generated by either manual curation of literature-derivered evidences or a dedicated pipeline and stored in locally managed MongoDB instances. Upon user interaction, the client layer sends HTTP requests to an underlying software layer that is able to interact with a database instance. When the database operations are completed, the results of the queries are sent back to the client layer in JSON format and get displayed in a standard web browser.

protein data such as name, sequence, organism of origin and a list of alternative names. Moreover, three lists of cross-references are available for each entry. The first is a list of PDB identifiers representing 3D structures for the protein, the second is a list of Pfam domain identifiers and the third is a list of annotations from the previous release MobiDB. Next to the core data collections, DisProt stores data about the biocuration team that generated the ID annotations. For this, a separate collection was deployed. This data organization is easy to maintain but very close to the relational paradigm: entries have a fixed schema with nullable fields. Implication of this schema is the need to query multiple collections to fetch data for a single protein. In the back-end, data aggregation routines had to be implemented in order to provide data for front-end layer and headless REST calls. Overall, this organization is very similar to relational DBMS with the extra complexity of implementing a the aforementioned "join routines".

MobiDB. The MobiDB database contains intrinsic protein disorder annotation for every protein from the UniProt database [230]. Annotations are organized in three tiers reflecting the confidence in information provided by the source. Internally, MongoDB documents schema reflects this organization even if a single collection holds data for all proteins. Here, each protein is represented by a document. Within each document, along with basic data about the protein (protein name, synonims, organism, sequence etc.) variable fields represent the available layers of data: the `mobidb_consensus` field reports aggregated data from the three layers, `mobidb_data` field lists all available data from the three data layers, finally special fields `pdb` and `bmrb` report cross-references and metadata from these two specialized databases. This internal organization conforms with the non-relational phylosophy. Documents in the database do not have fixed structure and are completely unrelated one to the other. This approach further simplifies database management.

4.2.2 REST back-end

In both DisProt and MobiDB, the back-end service is implemented using the Node.js framework. DisProt back-end is implemented using Node.js 0.10, while MobiDB using Node.js 6.9.0. These services are implemented as RESTful API: interaction with the data is granted via standard HTTP requests using specific URL addresses. Query parameters are either specified as query strings in URL of GET requests or in the body of POST requests. Both services return JSON documnents. Advanced users can use these bypassing graphical user interfaces as the endpoints are publicly accessible. This functionality relies on the JS Restify library 3.0.2 for DisProt and 5.0.1 for MobiDB.

Both web-services use JSON as primary data format. This choice simplifies the interaction with underlying MongoDB DBMS by matching its internal data representation model. Utilization of Node.js further simplifies database interaction: the JSON format is directly derived from JS objects representation, thus, using a JS-based framework allows direct and transparent conversion between these two data structures.

4.2.3 Front-end

Front-end layers for the two applications rely on HTTP endpoints exposed by their respective back-ends. Using data returned by these calls, they process and visualize them through a standard web-browser. These layers are developed using different versions of the same JS framework: for DisProt, the front-end is implemented using Angular.js 1.4.8 while for MobiDB Angular 4.4.4 [231] has been used. Version differences imply a large amount changes in programming patterns. The main one is the transition from vanilla JS (Angular.js) to Typescript, a syntactical superset of JS that introduces static typing. The main advantage of using statically typed languages is that they help programming tools, namely integrated development environments and compilers, to understand code while writing it, allow the usage of tools for syntax checking and static code analysis reducing bugs, syntactical errors and segmentation faults. For both projects, the development of these layers was separate from others improving overall modularity and code reusability.

Albeit technical differences, the architecture of these layers is very similar. Both of them have an home page, an entry page and static pages to provide help and information about the projects. DisProt front-end has also a browse page to surf the content of the entire database with a tabular view. MobiDB, on the other hand, implements a custom search engine. Graphically, also intrinsic protein disorder annotations are depicted similarly: a widget, called features viewer, depicting a protein as horizontal linear axis and annotations on vertical axis has been forked from public space and adapted with advanced features to suit specific visualization needs for these resources. MobiDB implements also sequence and structure viewers highlighting annotations either on primary or tertialy protein structures.

5 RESULTS

5.1 Gene expression regulation via chromatin accessibility in prostate cancer progression (Publication I)

To characterize chromatin structure during PCa progression, ATAC-seq data were generated. Reads were aligned to the human genome (GRCh38), quantified and peaks detected. A custom procedure for signal normalization was implemented, the genome split in non overlapping windows and DARs detected. Identified features were used to investigate the role of chromatin accessibility in regulating gene expression during PCa progression.

5.1.1 Identification of genes candidate regulatory regions

In order to establish the relationship between chromatin accessibility and gene expression, correlation between ATAC-seq signal and gene expression values (RNA-seq and sRNA-seq) was computed. Both Pearson and Sperman correlation coefficients were calculated. To identify either proximal and distal regulatory sites, the analysis was run in different genomic contexts (Figure 4.2):

1. **Gene's TSS**. Globally, correlation coefficients show only moderate values (Spearman $= 0.11$, Pearson $= 0.11$). In PC to BPH comparison, enrichment of upregulated DEG with increased accessibility at their respective TSS is observed (Fisher exact test $p < 10^{-16}$). Several oncogenes related to PCa like *MYC*, *AR* and *BCL11A* follow this pattern. In PC to CRPC comparison, an enrichment of downregulated DEG with reduced chromatin accessibility in their respective TSS was observed (Fisher exact test $p = 9.19 \cdot 10^{-16}$).
2. In **gene's promoter area** 418 peaks, one BPH to PC DAR, and 9 PC to CRPC

Table 5.1 Number of genes with expression correlating with chromatin accessibility.

	Genes	DEG
TSS	713	99
Promoter DARs	10	2
Promoter peaks	414	68
DARs to closest gene	157	45
Peaks to closest gene	1,788	307
DARs in TAD	1,396	239
Peaks in TAD	8,697	861
Total available	20,008	1,424

DARs have strong correlation with expression of nearby gene. Eight out of nine PC to CRPC DARs show increased accessibility. The only closing one is located in the promoter of the *MIR30A* tumor suppressor miRNA [232] which shows also downregulation in CRPC ($\log_2$ fold-change = -1.24, Spearman ρ = 0.7 p = $6.7 \cdot 10^{-5}$).

3. **The closest genes to each ATAC-seq feature** shows strong association for 2265 peaks, 42 BPH to PC DARs and 136 PC to CRPC DARs. Distance between peaks and genes varies from 769 kbp upstream to 1.03 Mbp with a median of 602 bp and standard deviation of 105 kbp. For BPH to PC DARs it varies from 89 kbp upstream to 188 kbp downstream with a median of 4.2 kbp and a standard deviation of 51 kbp. Finally, for PC to CRPC DARs, distance varies from 348 kbp to 782 kbp with a median of 8 kbp and a standard deviation of 90 kbp.

4. **Genes and ATAC-seq features within the same TAD** mostly contribute to the detected correlation. This observation supports the idea that distal regulation controls changes in gene expression: 27,274 peak-gene pairs in 1,860 TADs and 3,535 DAR-gene pairs in 526 TADs were identified to have strong correlation. Peak-gene pairs consist of 17,066 unique peaks and 8,697 unique genes. DARs-gene pairs consist of 284 unique BPH to PC DARs and 1,037 PC to CRPC DARs. In other words 9.6% of all peaks, 16.4% of all BPH to PC DARs and 29.6% of PC to CRPC DARs were associated with at least one target gene within their TAD.

Overall, from the joint analysis of gene expression and proximal accessible features (analyses 1 and 2) a weak global correlation emerges. In spite of this, the change in expression of some disease-relevant genes is strongly associated with change of chromatin accessibility suggesting promoter reconfiguration during transitions from benign hyperplasia, primary PCa and after androgen deprivation therapy.

From analyses 3 and 4 the role of distal regulation in driving cancer progression emerges. The significant detectable correlation between accessibility signal and gene expression suggests that these regions are actually involved in regulation of target genes and may function as enhancers.

Taken together, 45.5% of genes were associated with at least one ATAC-seq feature. The expression of 42 known oncogenes (e.g. *EGFR*, *ERBB2*, *JUN*, *FGFR1* and *FGFR2*), 27 tumor suppressor genes (e.g. *NOTCH1*, *BRCA1*, *BRCA2*, *IL2*) and 22 chromatin remodellers (e.g. *HDAC1*, *HDAC2*, *HDAC5*, *HDAC6*, *HDAC9*, *HDAC10*, *SMARCD1*) correlated with the chromatin accessibility of at least one peak. Furhtermore, the expression of 4 oncogenes (*JUN*, *PIM1*, *CARD11* and *TFG*), 5 tumor suppressor genes (*PTEN*, *NOTCH1*, *CDK6*, *FH*, *WT1*) and 2 chromatin remodeling factors (*HDAC7* and *CHRAC1*) were strongly associated with DARs accessibility.

Genes with differential expression pattern, show stronger correlation with chromatin accessibility: 62.4% of them show correlation with at least one ATAC-seq feature. This supports the idea that majority of progression-related gene expression variation can be explained by chromatin accessibility in PCa progression. Chromatin-accessible features associated with gene expression are mostly located upstream of TSSs (median distance 5.6 kbp). Moreover, 44.2% (4,051) of the genes with expression correlated to chromatin accessibility are linked to exactly one regulatory element and 107 genes (1.2%) can be associated to 30 or more regulatory elements (mean = 1.7). Similarly 71.6% (13,983) of peaks or DARs correlating with gene expression are associated with a single gene and 48 are linked to 30 or more unique genes, indicating that those might be super enhancers (mean = 0.2).

At a global scale, the results of these analyses show that the degree of correlation is not detectable: median correlations values are Pearson $p = 0.04$ and Spearman $\rho = 0.03$. Some ATAC-seq features, however, show strong correlation values (Pearson $p \in [-0.8, 0.92]$, Spearman $\rho \in [-0.79, 0.85]$) implying that chromatin structure is relevant in regulation of key genes involved in the disease.

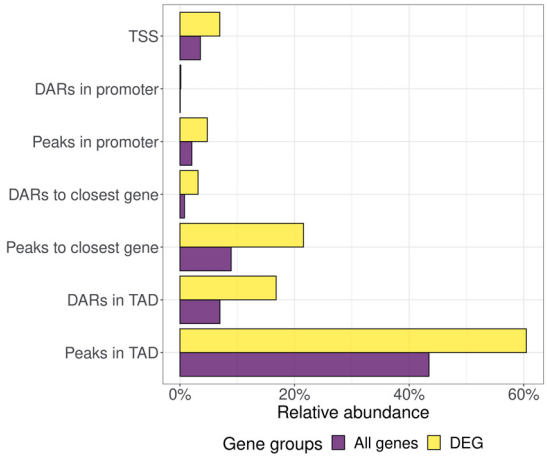


Figure 5.1 Relative abundance of significant correlations detected in different genomic contexts. Raw values are reported in Table 5.1.

5.1.2 Identification of transcriptional programs involved in prostate cancer progression

Candidate regulatory regions were identified from previous analyses. These regions were selected to display a strong correlation between Tn5 accessibility and expression of target gene.

In order to detect TFs governing the expression of genes driving PCa progression, a regulative network was generated. It connected TF sharing target genes via DARs correlating with gene expression and displaying binding sites. Groups of TF sharing target genes were identified and defined as regulative modules.

The module connected to most genes included AR, FOXA1, ESR1, and ERG. These TF are well known for their involvement in sustaining tumor growth. Another module included, among the others, SP1, HOXB13, TP63 and FLI1 (Figure 5.2). The genes connected to these latter TFs were a subset of those connected to the members of the AR module.

Shared binding sites for PCa-specific TF were searched in accessible features to study their co-localization pattern. Regions of interest were split by increased or reduce chromatin accessibility and proportions of detected binding sites compared. AR was the most frequent TF in all conditions in both comparisons: its binding sites were present in 92% (215) and 48% (24) of opening and closing BPH to PC DARs.

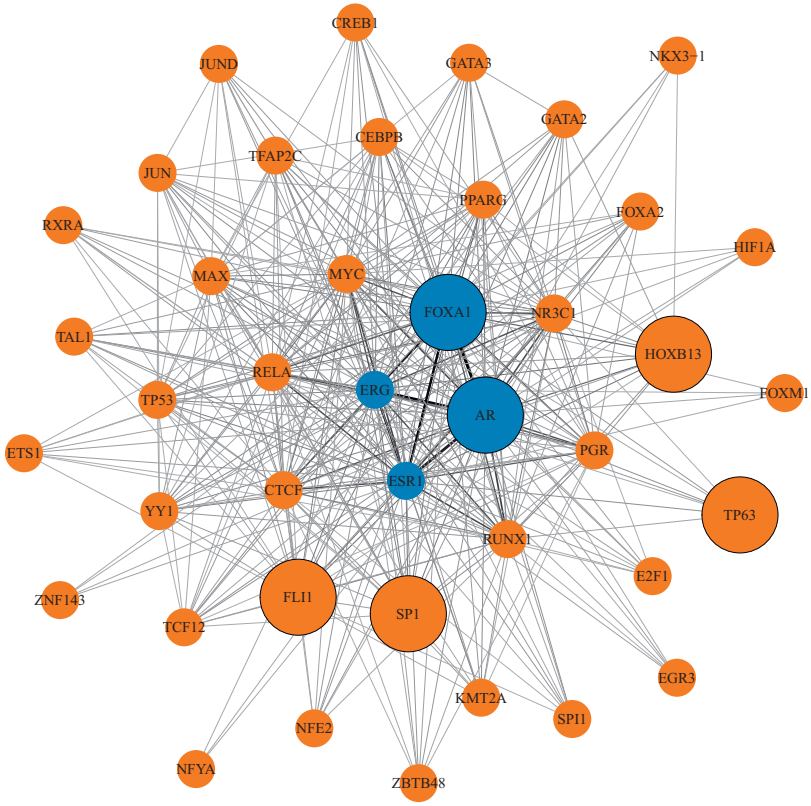


Figure 5.2 Regulative network of TFs involved in PCa progression. Target genes were assigned to each TFs via the correlation analysis described in Section 5.1. From the network emerges a core AR module and the secondary one of other factors. TFs described in Section 5.4 are highlighted by larger, circled nodes. Edges are drawn between couples of TFs that share more than 278 genes (55% of the distribution of shared genes). Edge color and width are proportional to the number of shared genes: wider, darker edges represent higher number of shared genes.

Among these, 58.1% of opening and 4.3% of closing DARs, reported binding sites also for FOXA1 and HOXB13. This observation hints to increased availability of TF binding sites for factors belonging to the AR module during transition to primary PCa. On the other hand, in PC to CRPC comparison 25% (209) and 92% (177) of opening and closing DARs, respectively, reported AR binding site. Of these, 6.4% of the opening ones displayed FOXA1 and HOXB13 binding sites, while 39.2% of the closing ones report this pattern.

To investigate the role of TFs from the other module, the analysis was expanded to all available TF. Opening DARs show augmented number of TF binding sites for SP1, TP63 and FLI1 with respect to closing PC to CRPC DARs. This suggests that, albeit sites available in primary PCa are target of AR and mostly stay open over the transition to advance PCa, new sites open during the progression enabling other TFs to bind and complement canonical AR regulative action. These TF, thus, act as co-modulators of the AR-driven stimulation of gene expression contributing to the more aggressive castration resistant phenotype observed in later stages of the disease.

5.2 DisProt (Publication II)

5.2.1 Database description

DisProt collects high quality literature-derived manually-curated annotations on IDPs. Two types of documents are stored in the database: "protein" and "disordered region" (IDR) documents. The former collects general information on a protein, while the latter holds disorder evidences from literature. Nomenclature of both proteins and IDR have been standardized with respect to previous database releases.

Protein information is taken from UniProt and stored in DisProt in a sequence-centric manner. Proteins isoform are stored separately because isoforms are produced by different mRNA. On the other hand cleaved proteins are merged in unique entries because they originate from the same precursor chain.

Disordered regions information is stored in an evidence-centric manner. For each experiment on a region, separate documents are stored. This pattern allows for simpler tracking of literature reference and for the introduction of a system to flag inconsistencies. DisProt reports three types of inconsistency: ambiguous sequence, ambiguous literature and ambiguous experiment. The fist type aims at reporting disorder evidence determined in non-physiological sequences, e.g. engineered sequence or fragments. The second is used to describe unclear statements about regions and the third issued when experimental conditions used to determine the disorder state of the region are too different from physiological state. Each region carries also functional annotations. The set of controlled terms used to describe IDR functions were generated during the annotation process by expert biocurators that reviewed literature and generated the data. These terms are organized in three separate sets aiming to

Table 5.2 DisProt organisms. Organisms with less than 5 annotated proteins were aggregated.

Organism	Number of IDP
Homo sapiens	231
Saccharomyces cerevisiae	49
Escherichia coli	44
Mus musculus	44
Bos taurus	31
Rattus norvegicus	30
Arabidopsis thaliana	20
Drosophila melanogaster	13
Human immunodeficiency virus type 1 group M subtype B	9
Escherichia coli O157:H7	8
Gallus gallus	7
Bacillus subtilis	6
Escherichia coli	6
Mycobacterium tuberculosis	6
Glycine max	5
Methanocaldococcus jannaschii	5
Oryctolagus cuniculus	5
Spinacia oleracea	5
Sus scrofa	5
Others	227
Total	756

describe different functional aspects of IDR by standardizing the terminology.

Proteins collected in DisProt 7 come from 198 different organisms (Table 5.2) and have been studied using 36 different experimental procedures (Table 5.3).

5.2.2 Disorder functional ontology

The term "intrinsically disordered protein" is used to describe a continuum of structural states a protein can take. These range from folded to fully unfolded. The

Table 5.3 Experimental methods used to identify IDRs in DisProt. Methods with less than 10 annotated IDR have been aggregated.

Method	Number of IDR
Nuclear magnetic resonance	1,434
Missing electron density	1,329
Circular dichroism spectroscopy far-UV	932
Sensitivity to proteolysis	239
Proton-based nuclear magnetic resonance	198
Size exclusion/gel filtration chromatography	179
Circular dichroism spectroscopy near-UV	102
Sodium dodecyl sulfate polyacrylamide gel electrophoresis	93
Fourier transform infrared spectroscopy	82
Small-angle X-ray scattering	78
Dynamic light scattering	64
NMR-based hydrogen-deuterium exchange	61
Analytical ultracentrifugation	54
Fluorescence intrinsic	51
Immunochemistry	29
Stability at thermal extremes	28
Differential scanning calorimetry	26
Fluorescence polarization/anisotropy	24
Site-directed spin-labelling electron paramagnetic resonance spectroscopy	20
Rotary shadowing electron microscopy	19
High relative B-factor	14
Mass spectrometry-based high resolution hydrogen-deuterium exchange	13
Atomic force microscopy	11
Electrospray ionization fourier transform ion cyclotron resonance mass spectrometry	11
Viscometry	10
Others	61
Total	5,162

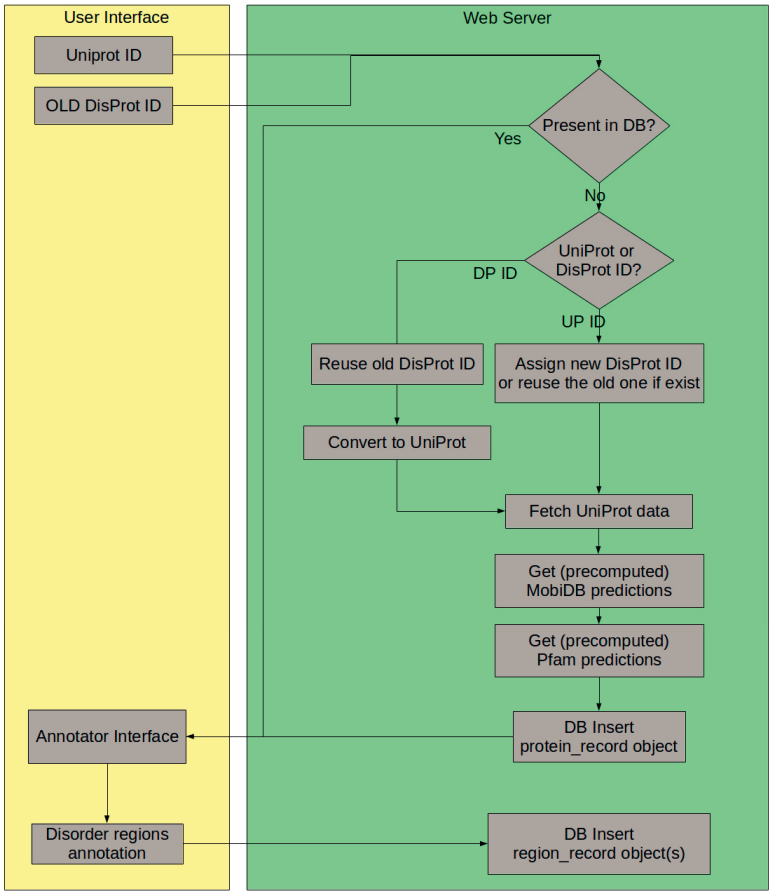
structural state of IDP and IDR has functional implications: IDP in a structured state may have different functions than in a unstructured state, thus it is important to categorize functional implications a structural state can have.

The DisProt ontology is inspired from Gene Ontology [233, 234], the most widely used, standardized resource for description of biological molecular species. The Gene Ontology is organized in three branches: biological process, molecular function and cellular component. Altogether these three branches can be used to categorize genic product with a standardized terminology. Similarly, the DisProt ontology, is divided into three branches that describe three different functional aspects of IDPs. The main focus of these terms is on the molecular function because it is the most different from globular proteins and enzymes. Other two branches aim at describing molecular structural transition an IDP undergoes in order exert its function and the interaction partner it binds to perform it. The DisProt ontology provides basis for standardized description of IDPs functional aspects.

5.2.3 Biocurator interface

As mentioned above, DisProt data has been manually annotated by a team of expert curators. To achieve this, they used a dedicated publicly available interface specifically developed for this purpose. Upon login, a biocurator could search a dedicated temporary collection to retrieve older entries or generate new ones. In either case, protein data from UniProt [230, 235], disorder predictions from previous MobiDB release [154] and Pfam annotations [236] were collected and stored in the DisProt database (Figure 5.3A). Next, the curator could add data from literature to describe a disorder region and store it remotely using the aforementioned graphical interface (Figure 5.3B). To generate a new annotation, biocurators had to insert intrinsically disordered region coordinates on a protein chain together with a PubMed id of a paper reporting the experimental identification. On top of these mandatory information, curators could report experimental detection methods, cross-reference PDB [237, 238], add functional annotation, name the region and add a free text comment.

A



B

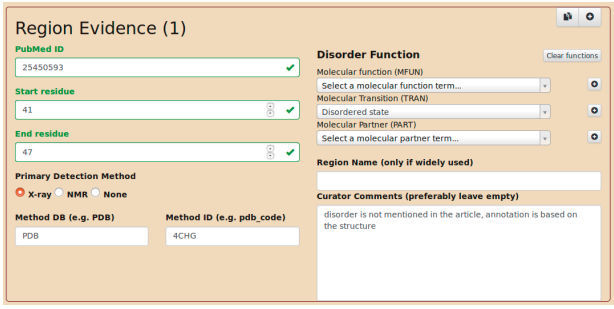


Figure 5.3 DisProt biocurator interface. **A.** Graphical representation of the algorithm used to prefetch data upon biocurator request to generate a new entry or update an existing one. **B.** The web form used by biocurators to upload literature-derived annotations on IDRs for an hypothetical protein.

5.2.4 Data accessibility

DisProt data are accessible through the web-interface and a REST application programming interface (API) (Figure 4.3). The web-interface featured a tabular view that allows for quick subset and search of entries. This view allowed to download the entire dataset or a subset generated by filtering the table. Multiple download formats were available.

For each protein, an entry page was available. At its top it showed general information from UniProt, followed by a summary view of all annotated IDRs and, finally, a detailed list reporting all IDR annotations. For each region, coordinates on the protein, experimental identification procedure, literature reference and name of the curator are reported.

Advanced users could take advantage of the REST API to directly query the database and download custom subsets of the data or extend third-party applications by including DisProt annotation in other custom views.

5.3 MobiDB (Publication III)

5.3.1 Database description

MobiDB represents the central repository of intrinsic protein disorder annotation because it provides intrinsic protein disorder annotation for the entire protein universe. Data in the database is organized in three tiers: the top one is made of manually curated data from external databases, the second is derived from indirect experimental evidences and the third is made of predictions. For each protein, three annotation types are available: direct evidences of ID, LIPs and secondary structure population. These two complementary aspects shape documents in MobiDB. Annotations displayed in the web interface are a synthesis of all available information computed by prioritizing curated and indirected evidences but detailed data are visualized as well.

Curated annotations. Curated annotations are collected from ten databases. UniProt and DisProt provide general curated disorder annotation. DisProt annotation was propagated to homologous proteins using GeneTree [239] alignments. FuzDB [240] provides data for regions involved in fuzzy complexes which are relevant for pro-

tein complex formation, regulation and higher-order assemblies. ELM [241] and IDEAL [242, 243] provide curated annotation for LIPs. The term "LIPs" is vague in litearature: in ELM, these features are reported as "SLiMs" (short linear interacting motifs), in previous literature they are referred as "MoREs" (molecular recognition features), while in IDEAL they are called "ProS" ("protean" segments"). In MobiDB, the term linear intercting peptide (LIP) is used to describe such features from all these different sources aiming at unifing the terminology. More annotations are pulled from two specialized databases Disordered Binding Site database (DIBS) [244] and Mutually Folding Induced by Binding database (MFIB) [245]. The former reports annotation for folding upon binding proteins, while the latter for mutually induced folding regions. Gene3D [246] provides curated annotations and prediction on protein structure, complementing MobiDB annotation. Curated and predicted Pfam [236] structural domains are also used. Moreover, CoDNaS [247, 248] provides indirect annotations for conformational diversity from NMR experiments.

Indirect annotations. Indirect annotations are derived from PDB [249, 250] and BioMagResBank (BMRB) [251]. Disorder information can be inferred from X-ray crystal structures deposited at PDB by analyzing three parameters: high temperature, missing residues and mobile residues. These analyses are carried out using the MOBI 2.0 [155] and RING [167] servers. High-temperature is calculated from B-factor regions, while missing residues are inferred from X-ray and cryo-EM structures by comparing the experimental PDB sequence with the observed crystallized one. Mobile residues are calculated using structural ensemble from NMR experiments by comparing the displacement of protein backbone in different aligned models. Also LIP can be inferred from molecular structures and are calculated by comparing inter- and intra-chain contacts. In this context, they are defined as regions with double number of inter-chain than intra-chain contacts. Chemical shifts are measurement of fluctuations in protein structures and BMRB is a publicly available repository for such data. NMR experiments allow for measurement of chemical shifts and their analysis allows the reconstruction of ensemble of secondary structure elements transiently forming in a peptide chain. This data is useful to identify protein regions that fold for a short period of time (up to the millisecond scale) and help to overcome the order/disorder dicotomy. Together with secondary structure populations, MobiDB reports experimental conditions at which these transient folding was observed. If

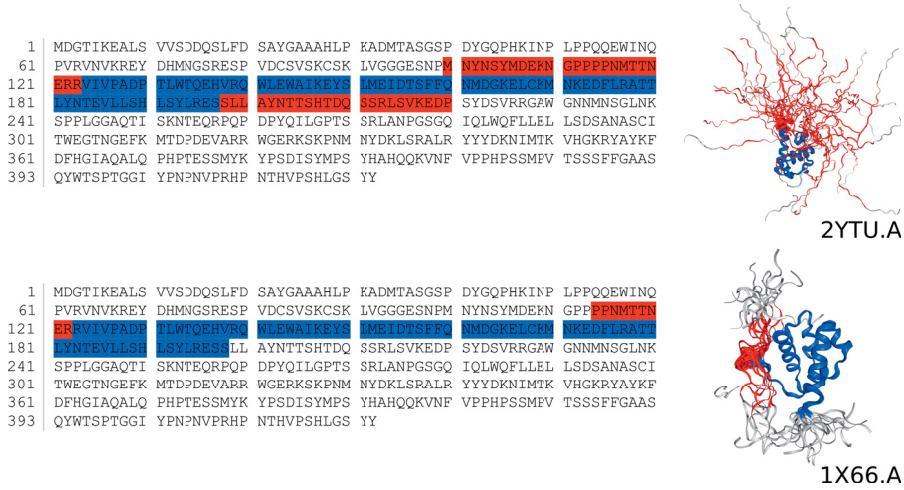


Figure 5.4 Visualization of the annotations from MobiDB 3.0 on protein sequence and structure. In red, IDRs as computed from mobile residues on the two reported PDB structures. In blue, residues sitting in a folded region on the protein. Entry: FLI1, UniProt: Q01543.

multiple chemical shifts were available, an overview was calculated and presented to the user.

Predictions. Intrinsically disordered protein predictions are computed using ESpritz [165], IUpred [160, 161, 162], DisEMBL [169] and VSL2b [171]. DynaMine [175] is used to predict backbone flexibility, Anchor [164] to predict binding sites within disordered regions and LIPs. Secondary structure elements are predicted with FELLS [166], δ2D [156] and RCI [157, 158, 159]. Separate predictions are merged in a consensus using the MobiDB-lite tool [252] which implements a majority vote strategy to unify separate predictions and enforces 20 consecutive residues to report a disordered region [252].

5.3.2 Data accessibility and visualizations

Similarly to DisProt, data stored in MobiDB are accessible via standard web browser and using a REST application programming interface (API) (Figure 4.3). The web interface, for each entry, all available data are reported with a specialized interactive feature viewer. Data are organized in subsets and the user can select which to display. Four groups are defined: curated, indirect, predictions and interaction. Data in the feature viewer are organized in vertical tracks with the protein sequence acting as

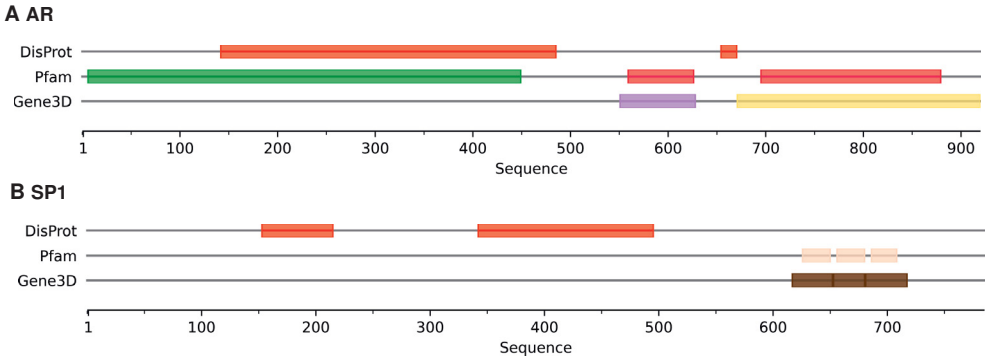


Figure 5.5 DisProt annotations for TFs involved in PCa progression. **A. AR** - DisProt: DP00492. UniProt: P10275. Pfam: zinc finger nuclear hormone receptor-type (PF00105), nuclear hormone receptor ligand-binding domain (PF00104). Gene3D: erythroid transcription factor GATA-1, subunit A (3.30.50.10), retinoid X receptor (1.10.565.10). **B. SP1** - DisProt: DP00378. UniProt: P08047. Pfam: Zinc finger C2H2-type (3x PF00096). Gene3D: classic Zinc Finger (3.30.160.60). Pfam and Gene3D domains are listed from N- to C- terminal. Color codes are the same as Figure 5.6.

horizontal axis. Each feature, in each track, is color coded to simplify interpretation (Figure 5.5 and Figure 5.6). By selecting any feature, textual information are displayed. Data contributing to a specific feature track can be shown by clicking on the track label. Features derived from molecular structures can also be viewed in an interactive three-dimensional viewer that depicts the structure maintaining the color code of the main viewer (Figure 5.4). Same features can be displayed also on the primary sequence with a dedicated widget that behaves similarly to the three-dimensional viewer (Figure 5.4).

The MobiDB API can be used by advanced users to interact with the database and retrieve specific subsets or organism-specific data and perform custom queries. The web-interface provides access a set of pre-computed datasets and implements a search engine which allows the user to browse the database.

5.4 Structural features of transcription factors involved in primary prostate cancer progression

Transcription factors from the modules described in Section 5.1.2 share structural features. Structural annotations for all these proteins are available in PDB and Pfam.

Intrinsic protein disorder annotations for AR and SP1 can be found in DisProt, while ID annotations for FOXA1, HOXB13, FLI1 and TP63 can be retrieved from MobiDB.

AR The AR TF is a nuclear steroid receptor. Pfam reports three structural domain: an N-terminal modulatory domain, a C-terminal hormone binding domain and a DNA binding domain (Figure 5.5A). Receptor activation upon hormone binding on the C-terminal domain, induces the folding of a helix disclosing a groove that binds to the N-terminal modulatory domain. This displacement allows the protein to dimerize and to bind its target sequence on the DNA. The DNA binding domain is composed of two zinc-fingers: the first binds to the DNA major groove, while the second stabilizes this interaction and is required for receptor dimerization. The folding-upon-binding event that characterize AR activation is reported in DisProt by three experiments. These, characterize a region spanning 342 residues with different techniques. The three sources agree on the transition induced by the ligand and report it as a "disorder to order transition". They also add that the region is involved in protein binding and in modulation of the activity of a partner molecule, the hormone. To describe the role in dimerization, this region is annotated to be involved in the assembly of molecular complexes.

SP1 Pfam annotation for SP1 reports only the DNA binding domain which is made of three zinc finger modules. DisProt complements this annotation reporting data from nine experiments on two distinct IDRs. The first region is 63 amino acids long, from Pro 153 to Ile 215. The second IDR is 153 amino acids long, from Thr 342 and Thr 495. These regions correspond to binding sites detected to interact with TAF4 and to dimerize (Figure 5.5B).

FOXA1 For FOXA1 (Figure 5.6A), three Pfam domains are reported: the forkhead N-terminal region, the forkhead domain and a C-terminal region homologue to the Hepatocyte Nuclear Factor 3 (HNF3) α and β chains. Within this last domain, MobiDB reports a linear motif from ELM. No indirect ID annotation from X-ray or NMR data is reported from PDB. Predictions from MobiDB-lite report an IDR from Glu 269 to Pro 392, in a region between the forkhead domain and the C-terminal domain. Inspecting results from single predictors, heterogeneous predictions can be retrieved. Although such diversity, MobiDB-lite consensus prediction is backed by

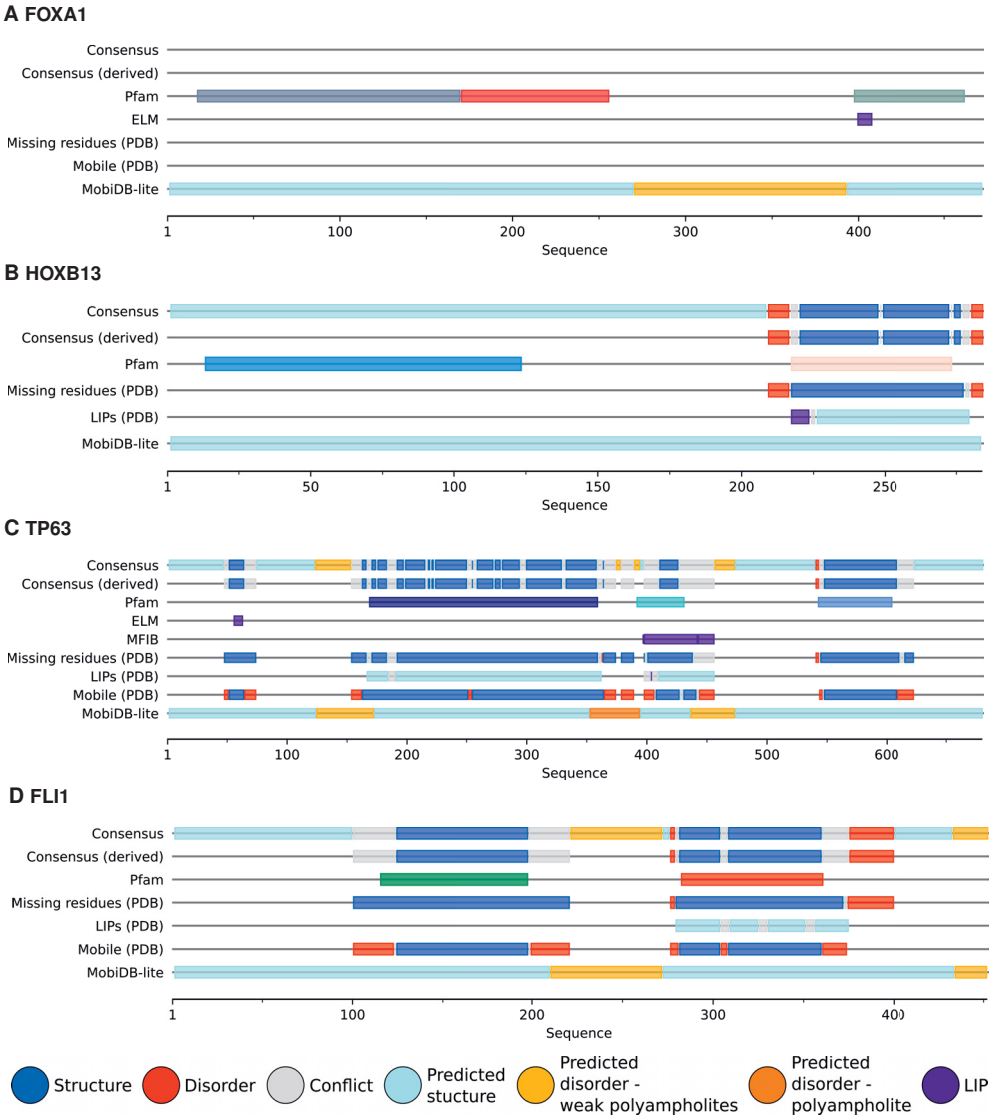


Figure 5.6 Summary of available annotations for TFs involved in PCa progression. **A. FOXA1** - UniProt: P55317. Pfam (N-terminal to C-terminal): Fork-head N-terminal (PF08430), Fork head domain (PF00250), Forkhead box protein C-terminal (PF09354). **B. HOXB13** - UniProt: Q92826. Pfam: homeobox protein Hox1A3 N-terminal (PF12284), homeobox domain (PF00046). **C. TP63** - UniProt: Q9H3D4. Pfam: p53 DNA-binding domain (PF00870), p53 tetramerisation domain (PF07710), sterile alpha motif domain (PF07647). **D. FLI1** - UniProt: Q01543. Pfam: Pointed domain (PF02198), Ets domain (PF00178). The "Consensus" track reports aggregated annotations from databases and predictions, "Consensus (derived)" aggregated annotations from databases only. Pfam, ELM and MFIB tracks report annotations from respective database. "Missing residues", "LIPs" and "Mobile" tracks report evidences from missing electron densities of X-ray-determined structures, linear intercting peptides and mobile residues in NMR structural ensembles deposited in PDB. Finally, "MobiDB-lite" reports the consensus annotation from all predictors run by the MobiDB-lite meta-predictor.

secondary structure predictions from FeSS and flexibility prediction from DynaMine: this region shows the lowest propensity to form helices or sheets, while predicted backbone flexibility is relatively high.

HOXB13 Two domains are annotated for HOXB13 (Figure 5.6B) from Pfam: A13 N-terminal domain and the homeobox domain. A portion toward the N-terminal of the protein has been studied with X-ray crystallography, thus, six crystal structure of this fragment are available from PDB. Altogether, they provide indirect evidence of a LIPs from Arg 217 to Tyr 223. From missing electron density, two short IDR can be inferred. The first one from Asp 209 to Gly 216 or Arg 217, depending on the PDB entry, and the other from Lys 277 to Pro 284, the last residue of the protein. These two regions are annotated around the homeobox domain. The ambiguity in region boundaries from missing electron densities (Gly 216 or Arg 217) is reported in the feature viewer as a conflict. For this protein, the consensus computed by MobiDB-lite reports a completely folded or structured protein. This result emerge from the disagreement of all underlying predictors.

TP63 For TP63 (Figure 5.6C), the N-terminal TP53 DNA-binding domain, TP53 tetramerization domain and a C-terminal SAM interaction domain are reported from Pfam. From ELM, a degron motif from TP53 N-terminal domain is retrieved. Curated annotations from the MFIB database report two features overlapping the tetramerization domain. These annotations refine the Pfam ones by reporting that these features are TP63/TP73 specific and add the information about mutually induced folding event. ELM and MFIB annotations are used in MobiDB to infer LIPs locations. For TP63, a large amount of indirect intrinsic protein disorder evidences is available. From X-ray resolved structures from PDB, a similar conclusion can be drawn. Because the crystallized protein is just a fragment of the whole peptide chain, only partial description of its structural organization can be drawn. Nevertheless, from inspection of the X-ray data, indirect evidences of a short LIP region overlapping the tetramerization domain can be inferred. Missing electron density in the same crystal from PDB reports a folded regions with ambiguous short missing residues. From mobile portions of structural ensemble derived from NMR experiments deposited at PDB, an IDR can be indirectly derived: five experiments contribute to build evidences of multiple short IDR overlapping the tetramerization domain. Evidences from chemical shifts complement these observations by providing more precise annotations of

secondary structure elements within the DNA-binding and tetramerization domains. Predictions from MobiDB-lite and secondary structure elements confirm the indirect evidences derived from databases. MobiDB-lite predicts three IDR: one overlapping with the DNA-binding domain, and the other two localized in the neighborhood of the tetramerization domain. Prediction of secondary structure elements from FeSS, confirm these predictions showing lowest prediction scores for helix and sheets.

FLI1 The *FLI1* gene codes for an ETS TF. Pfam annotation coherently reports (Figure 5.6D) the ETS DNA-binding domain toward the C-terminus of the protein. The sequence of the domain confers these TFs their binding specificity. Like other proteins of the same family, FLI1 shows a highly conserved Pointed domain toward the N-terminus which is used for protein interaction. Indirect evidences from databases report IDR flanking these two domains. LIPs annotation derived from PDB structures is ambiguous and depends on the stucture under analysis. Missing electron densities agree on a IDR downstream of the ETS domain. Residues on the flanking region of ETS domain are relevant for binding specificity of the transcription factor. Mobility data from PDB extend this data by adding two IDRs on flanking regions of the Pointed domain. Predictors report slightly different results: the MobiDB-lite consensus reports one long region between the two domains. Inspecting predictions from underlying tools, seven out of ten tools report an IDR close to the Pointed domain, while five out of ten report another IDR downstream of the ETS domain. Secondary structure predictions from FeSS report highest scores in regions annotated with the mentioned structural domains. IDR annotation inferred either from indirect evidences or predictions agree on their location outside structural domains.

6 DISCUSSION

Prostate cancer growth is sustained by testosterone stimulation. After first line treatment, primary PCa frequently relapses to a more aggressive, testosterone insensitive form. Benign prostate hyperplasia is characterized by localized hypertrophy of the prostate gland. Primary PCa differs from BPH for the elevate PSA secretion and sensitivity to androgen stimulation. According to detected PSA level, treatment options vary: watchful waiting, surgery, radiotherapy or hormonal treatment can be used to treat primary PCa in the clinic. The relapsed stage, castration resistant prostate cancer, can be treated with antiandrogen therapies.

Primary PCa is characterized by widespread genomic alterations, e.g. fusions, amplifications, deletions and SNPs on chromosomes 7, 8, 13, 17 and X involving loci conding for oncogenes and tumor suppressor genes such as *ar*, *MYC*, *ERG*, other ETS genes and HOX genes. Transition to CRPC further selects for highly aggressive and resistant phenotypes. Typical mutations of this stages are deletions of chromosome 17q that hosts *TP53*, *BCL2* overexpression and abnormal activation of the PI3K/Akt pathway.

Chromatin structure and gene expression are two fundamental processes altered during cancer progression to CRPC. Chromatin structure regulates gene expression by allowing or denying physical access to binding sites to TF and by acting on the speed RNA polymerase II synthesize pre-mRNA [253, 254].

6.1 Role of enhancers in prostate cancer progression

To investigate how chromatin structure interact with gene expression, chromatin accessibility across the three PCa stages was quantified with ATAC-seq. For the same samples, methylation, RNA-seq and proteomics data were available.

Gene expression is a tightly regulated process, involving coordinated interaction of many biochemically heterogeneous partners. First, responsive element recognition

on genomic DNA involves protein-DNA interaction. This process requires that the target genomic sequence is physically available for binding. Availability is determined by factors like nucleosome positioning and overall chromatin compression. Next, to activate gene expression, the transcription machinery must be assembled and activated. This process requires interaction of multiple DNA bound factors and cofactors. Extensive genomic DNA structural rearrangements and phase-separation drive these interactions. In pathological conditions, this complex and finely regulated mechanism gets altered. Such alterations result in cell cycle perturbation, tumor mass development, subclonal evolution, migration and metastasis.

To elucidate the relationship between chromatin accessibility and gene expression, correlation was calculated among ATAC-seq and RNA-seq in four different genomic context. First, proximal accessible features were considered. Overall, weak correlation between accessiblity of proximal elements and gene expression was detected. This observation confirms previous results [255, 256, 257, 258] and hints at the role of enhancers in gene expression regulations. To investigate this hypothesis, distal ATAC-seq features were correlated with gene expression. Closest gene to any feature and all pairs of genes and ATAC-seq features within the same TADs were also considered. Detected global correlation was weak but a stronger relationship was observed for a subset of DEG genes across disease stages.

Globally, through correlation analysis at least one regulatory region was assigned to 84.7% genes with differential expression. The majority of these assignments were from distal elements, highlighting the importance of enhancers in regulating gene expression during PCa progression. The reduced sample size, however, may hinder observations by introducing false positives: for example, strong correlations may be detected due to one or few outlier samples. Moreover, in an effort to minimize this phenomenon, too stringent threshold values might have been used resulting in an increased false negative rate. Furthermore, TAD coordinates used to detect genes and ATAC-seq features couples were determined in cellular models of PCa. These may not describe the actual state of chromatin in the sample cohort resulting in erratic associations between genes and ATAC-seq features. Another limitation of this approach is to compute only direct correlations. Complex patterns of interaction involving more than one enhancer with a single promoter, multiple enhancers or multiple promoters have been shown to have an additive effect on gene expression [24, 259, 260, 261].

6.2 Transcription factors involved in prostate cancer progression

Candidate regulatory regions were assigned to target genes through correlation analysis. Next, to elucidate transcriptional programs active in PCa progression, TF with binding potential to these regions were detected. Unsurprisingly, the one with the highest count of binding sites across all conditions and comparisons is AR. Similarly, on chromosome Xq12, expression of AR displays correlation with 48 peaks and 5 DARs from the PC to CRPC comparison, one of which overlaps with a previously reported enhancer [262]. These DARs display binding sites for AR, FOXA1, HOXB13 and ERG. These observations confirm the role of AR and other members of this module in PCa establishment and progression.

AR shares high number of putatively regulated genes with FOXA1 and HOXB13 [263]. Co-localizing binding sites for these three TFs were observed in DARs with increasing availability from BPH to PC comparison and reduced availability in PC to CRPC. This observation supports the idea that collaborative function of these proteins is required for primary PCa establishment, while for castration resistance, cooperative action of these factors becomes less prominent in favor of other co-modulating TFs. Co-modulators do not replace AR but rather collaborate to drive androgen independent tumor growth as demostrated by the reduced number of co-occuring binding sites for factors from the AR module in newly accessible features from PC to CRPC comparison. In section 5.1.2 three more co-modulating TFs were described: SP1, TP63 and FLI1.

The *FOXA1* oncogene codes for a transcription factor belonging to the forkhead protein family. It is known to have pioneering activity by recruiting bromodomain–containing proteins and histone methyltrasferases. These proteins increase local H3K27ac and H3K4me1/me2/me3 that, in turn, recruit other chromatin remodeling proteins to reposition histones. Augmented accessibility allows binding of AR to enhancer elements, interaction with the initiation complex assembling on the promoter of a target gene and activation of the transcription machinery [264].

The *HOXB13* tumor suppressor gene codes for a protein belonging to the homeobox family which is involved in cell differentiation during development. Specifically, HOXB13 belongs to the AbdB subfamily which is involved in differentiation of

posterior domain development including urogenital tract. HOXB13 interacts with AR [265]. The AR–HOXB13 complex represses expression of genes with androgen responsive elements alone and stimulates genes with both androgen and HOX responsive elements in their promoter [265]. It has also been shown that the HOXB13 germline variant G84E is a susceptibility factor for PCa development [266].

The *SP1* gene codes for an ubiquitous TF interacting with AR [267, 268], while *p63* codes for a protein member of the TP53 family. *FLI1* codes for an ETS family factor similar to ERG. Inhibition of SP1 inhibits tumor growth in cellular models [269].

Moreover, other transcription factors from the regulative network (Figure 5.2) are involved in PCa-related processes. For instance, RUNX1 together with RUNX2 is known to be involved in metastasis formation [270]. Moreover, RUNX1 has been show to be a downstream target of AR signaling and to have different functions in primary PCa and CRPC [271]. The expression of the prostate-restricted homeodomain containing TF NKX3-1 is sustained across different PCa stages and is anticorrelated with expression of MYC in murine models [272]. Similarly, expression of peroxisome proliferator-activated receptor gamma (PPARG) correlates with tumor grade and may be involved in progression [273]. Other TFs, e.g. MYC, MAX, TP53, have been known since early days of PCa biology research and their alterations have been since then associated with progression and aggressive phenotype [71]. Over-expression of the glucocorticoid receptor NR3C1 is associated with anti-androgen treatment and overall survival [274]. Finally, other TFs are associated with neuroendocrine phenotype, e.g. ERG, FOXA2 [275].

Altogether these findings show the relevance of these TFs in PCa progression and the feasibility of using complementary data types to investigate complex phenomena in the framework of cancer biology. Functional studies, though, are needed to confirm these results although *in silico* analyses provide candidates for functional validation and highlight the synergistic role of computational and experimental methods. From these analyses a set of candidate modulatory TF has been identified and has been associated with a consistent number of genes involved in PCa progression.

The detection of all the mentioned TF relied on the utilization of GTRD, a curated database reporting uniformly processed binding sites from ChIP-seq experiments. This choice limits the number of discoverable sites to the ones present in the database, precluding novel sites discovery. Although this limitation, the data is extremely valu-

able because of the uniform processing pipeline used to standardize them. Utilization of multiple peak calling algorithms boosts confidence on reported loci [229, 276]. On the other hand, all the data provided by GTRD come from cellular models which may not describe the DNA-protein interaction pattern of clinical samples under analysis.

In ATAC-seq features, often, multiple binding sites are observed. With the approach used for the analysis, all detected TFs are associated with a correlating gene. To understand which binding site is occupied and reduce the number of associations, a motif occupancy analysis could be run. In this case, the reduced sample size would have severely limited the confidence in results. Experimental functional validation would have been the ultimate proof of the predicted relationship between chromatin accessibility, transcription factor presence and gene expression regulation.

6.3 Structural features of transcription factors involved in primary prostate cancer progression

Intrinsic protein disorder is widespread in eukaryotes and the number IDP coded by a genome correlates with organism complexity. Intrinsically disordered proteins, mainly work as inter- and intra-molecular interaction hubs. The regulation of gene expression essentially works by allowing or denying molecular interaction among TFs, DNA and other proteins. It has been long acknowledged that TFs are a class of proteins with widespread intrinsic protein disorder content. For example, AR annotation from DisProt reports a large IDR involved in hormone binding, while SP1 annotation reports two IDR involved in dimerization and protein-protein interaction. These observations support molecular recognition function of IDPs. However, DisProt annotation is not available for the vast majority of proteins. The manual curation process required to add data both limits the amount of data that can be processed and ensure highest quality possible. However, albeit data quality, the curation process is not error-free and is prone to personal interpretation and misunderstanding. Nonetheless, because of data quality, DisProt serves as golden standard for training new prediction methods, benchmarking existing ones and providing detailed annotation on IDP.

MobiDB extends DisProt annotations to the entire protein universe. Proteins are

stored by their UniProt id, allowing direct linking of the two resourses. Annotations from UniProt, DisProt and a number of other specialized resources can be retrieved. Intrinsic protein disorder functional annotation is missing but, among identified TFs a pattern in the localization of IDR emerges: overlap with known structural domain is minimal. Transcription factors *in vivo* do not occupy all available binding sites, it has been shown that only a subset is actively used and that TFs scan the genome for their binding sites. Many speculations about the mechanism involved have been proposed and recently one has been reported to involve IDR with specific localization outside the DNA binding domain. Both the extent and sequence of these regions influence its binding specificity [2]. Intrinsic protein disorder, thus, can be linked both to protein-protein [277] and protein-DNA interactions [2, 148].

Altogether DisProt and MobiDB provide extensive annotations on intrinsic protein disorder. The IDP community relies on DisProt as the golden standard for high quality, manually curated data and on MobiDB as central repository for large scale protein annotation data. With the works presented in this thesis major updates to the two projects have been carried out and completely redesigned databases and web-applications have been deployed. Moreover, their utilization for *in silico* protein structure characterization has been presented. Currently, DisProt has been updated to version 8 by adding more annotations, improving the ID ontology and by deploying an updated web-interface [3] and soon a new major version of MobiDB will be released. Annotations provided by these two databases are helpful in the study of specific proteins from any biological process as they provide functional insights for regions involved in protein interactions and regulation. In the context of PCa these data can be used to characterize the structural features of protein regions relevant for cancer progression and establishment, guiding the design and interpretation of other experiments, drug design and, possibly, discovery of novel actionable targets.

7 CONCLUSION

Proximal regulative elements are required to localize the transcription machinery on genes promoters. On the other hand, distal *cis*-regulation is a fundamental mechanism to modulate transcription. Chromatin structure influences the ability of TFs to bind their regulatory sites. Similarly, TF binding may induce chromatin remodeling by recruiting specialized proteins. In pathological states, this mechanism is altered: chromatin gets remodeled resulting in increased accessibility of otherwise precluded TFs and vice-versa. Binding events on these loci modulate the expression of target genes in an abnormal manner. In cancer, these alterations are commonly observed. Because of this, characterization of the regulatory landscape of cancerous cells will have a clinical impact.

Prostate cancer growth is sustained by testosterone stimulation. After first line treatments, relapse happens in about a third of cases. CRPC cells become androgen-independent and can proliferate also in absence of hormonal stimulation. This switch can be explained by a selective sweep introduced by pharmacological treatments. Differential expression pattern can be largely explained by changes in chromatin accessibility: different genomic regions are involved in regulation of genes responsible for the transitions from BPH to PC finally to CRPC. The set of TFs binding these regions is restricted with AR being the most detected. Pioneering TF FOXA1 and HOXB13 are also found in many instances. Other TF can be detected in the two transitions with alternative patterns.

The structural organization of these proteins is substantially conserved. They share the ability to bind DNA via a DNA binding domain. To regulate such process all these factors display at least one IDR. These regions are characterized by high degree of flexibility and are often involved in molecular recognition and condensation driving the formation of phase-separated droplets in the nucleus. This process has been linked to PIC formation, transcriptional factories and chromatin remodeling complexes at super-enhancer loci. DisProt and MobiDB provide extensive annotations

for this phenomenon: they collect manually curated, database derived, indirect annotations and intrinsic protein disorder predictions from different resources and serve them through unified, standardized web interfaces. DisProt data is curated manually by a group of curators that also develops a controlled vocabulary of term to describe intrinsic protein disorder molecular function. MobiDB, on the other hand, brings together a multitude of other databases and tools. This extensive set of data is presented concisely and graphical tools are available to visually inspect them. Detailed data can be also retrieved from the online platform. Together these two resources consolidate the knowledge on intrinsic protein disorder and put the basis for greater underestanding of this property in all biological processes.

Altogether this dissertation provides insights into the regulatory landscape and of PCa demonstrating how the developed resources can be used to computationally understand and characterize the structure of detected proteins. This data-driven approach could be extended to the study of any disease or biological process. Transcriptomic and epigenetic data can be used to identify candidate regulators of the process and structural annotations databases can be used to characterize their three-dimensional structure. Common structural features can be further analyzed to find differences, e.g. by detecting common sequence motifs or used for antibody design, in drug design or repurposing.

REFERENCES

[1] J. A. Mitchell and P. Fraser. Transcription factories are nuclear subcompartments that remain in the absence of transcription. *Genes & development* 22 (1 Jan. 2008), 20–25. ISSN: 0890-9369. DOI: 10.1101/gad.454008. ppublish.

[2] S. Brodsky, T. Jana, K. Mittelman, M. Chapal, D. K. Kumar, M. Carmi and N. Barkai. Intrinsically Disordered Regions Direct Transcription Factor In Vivo Binding Specificity. *Molecular Cell* (June 2020). DOI: 10.1016/j.molcel.2020.05.032.

[3] A. Hatos, B. Hajdu-Soltész, A. M. Monzon, N. Palopoli, L. Álvarez, B. Aykac-Fas, C. Bassot, G. I. Benítez, M. Bevilacqua, A. Chasapi, L. Chemes, N. E. Davey, R. Davidović, A. K. Dunker, A. Elofsson, J. Gobeill, N. S. G. Foutel, G. Sudha, M. Guharoy, T. Horvath, V. Iglesias, A. V. Kajava, O. P. Kovacs, J. Lamb, M. Lambrughi, T. Lazar, J. Y. Leclercq, E. Leonardi, S. Macedo-Ribeiro, M. Macossay-Castillo, E. Maiani, J. A. Manso, C. Marino-Buslje, E. Martínez-Pérez, B. Mészáros, I. Mičetić, G. Minervini, N. Murvai, M. Necci, C. A. Ouzounis, M. Pajkos, L. Paladin, R. Pancsa, E. Papaleo, G. Parisi, E. Pasche, P. J. Barbosa Pereira, V. J. Promponas, J. Pujols, F. Quaglia, P. Ruch, M. Salvatore, E. Schad, B. Szabo, T. Szaniszló, S. Tamana, A. Tantos, N. Veljkovic, S. Ventura, W. Vranken, Z. Dosztányi, P. Tompa, S. C. E. Tosatto and D. Piovesan. DisProt: intrinsic protein disorder annotation in 2020. *Nucleic acids research* 48 (D1 Jan. 2020), D269–D276. ISSN: 1362-4962. DOI: 10.1093/nar/gkz975. ppublish.

[4] V. Haberle and A. Stark. Eukaryotic core promoters and the functional basis of transcription initiation. *Nature Reviews Molecular Cell Biology* 19.10 (2018), 621–637.

[5] G. Felsenfeld and M. Groudine. Controlling the double helix. *Nature* 421.6921 (2003), 448–453.

[6] K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent and T. J. Richmond. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* 389 (6648 Sept. 1997), 251–260. ISSN: 0028-0836. DOI: 10.1038/38444. ppublish.

[7] K. Maeshima, S. Hihara and M. Eltsov. Chromatin structure: does the 30-nm fibre exist in vivo?: *Current opinion in cell biology* 22 (3 June 2010), 291–297. ISSN: 1879-0410. DOI: 10.1016/j.ceb.2010.03.001. ppublish.

[8] D. Zhang, Z. Tang, H. Huang, G. Zhou, C. Cui, Y. Weng, W. Liu, S. Kim, S. Lee, M. Perez-Neut, J. Ding, D. Czyz, R. Hu, Z. Ye, M. He, Y. G. Zheng, H. A. Shuman, L. Dai, B. Ren, R. G. Roeder, L. Becker and Y. Zhao. Metabolic regulation of gene expression by histone lactylation. *Nature* 574 (7779 Oct. 2019), 575–580. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1678-1. ppublish.

[9] M. V. Liberti and J. W. Locasale. Histone Lactylation: A New Role for Glucose Metabolism. *Trends in biochemical sciences* 45 (3 Mar. 2020), 179–182. ISSN: 0968-0004. DOI: 10.1016/j.tibs.2019.12.004. ppublish.

[10] M. P. Creyghton, A. W. Cheng, G. G. Welstead, T. Kooistra, B. W. Carey, E. J. Steine, J. Hanna, M. A. Lodato, G. M. Frampton, P. A. Sharp, L. A. Boyer, R. A. Young and R. Jaenisch. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America* 107 (50 Dec. 2010), 21931–21936. ISSN: 1091-6490. DOI: 10.1073/pnas.1016071107. ppublish.

[11] G. E. Zentner, P. J. Tesar and P. C. Scacheri. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome research* 21 (8 Aug. 2011), 1273–1283. ISSN: 1549-5469. DOI: 10.1101/gr.122382.111. ppublish.

[12] K. Zhou, G. Gaullier and K. Luger. Nucleosome structure and dynamics are coming of age. *Nat. Struct. Mol. Biol.* 26.1 (Jan. 2019), 3–13.

[13] N. J. Francis, R. E. Kingston and C. L. Woodcock. Chromatin Compaction by a Polycomb Group Protein Complex. *Science* 306.5701 (2004), 1574–1577. ISSN: 0036-8075. DOI: 10.1126/science.1100576. eprint: https://science.sciencemag.org/content/306/5701/1574.full.pdf. URL: https://science.sciencemag.org/content/306/5701/1574.

[14] J. L. Goodier. Restricting retrotransposons: a review. *Mob DNA* 7 (2016), 16.

[15] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander and J. Dekker. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* 326.5950 (2009), 289–293. ISSN: 0036-8075. DOI: 10.1126/science.1181369. eprint: https://science.sciencemag.org/content/326/5950/289.full.pdf. URL: https://science.sciencemag.org/content/326/5950/289.

[16] K. P. Eagen. Principles of Chromosome Architecture Revealed by Hi-C. *Trends Biochem. Sci.* 43.6 (June 2018), 469–478.

[17] S. Saccani, S. Pantano and G. Natoli. Two waves of nuclear factor kappaB recruitment to target promoters. *J. Exp. Med.* 193.12 (June 2001), 1351–1359.

[19] B. R. Sabari, A. Dall'Agnese, A. Boija, I. A. Klein, E. L. Coffey, K. Shrinivas, B. J. Abraham, N. M. Hannett, A. V. Zamudio, J. C. Manteiga, C. H. Li, Y. E. Guo, D. S. Day, J. Schuijers, E. Vasile, S. Malik, D. Hnisz, T. I. Lee, I. I. Cisse, R. G. Roeder, P. A. Sharp, A. K. Chakraborty and R. A. Young. Coactivator condensation at super-enhancers links phase separation and gene control. *Science (New York, N.Y.)* 361 (6400 July 2018). ISSN: 1095-9203. DOI: 10.1126/science.aar3958. ppublish.

[20] L. A. Lettice, S. J. H. Heaney, L. A. Purdie, L. Li, P. de Beer, B. A. Oostra, D. Goode, G. Elgar, R. E. Hill and E. de Graaff. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human molecular genetics* 12 (14 July 2003), 1725–1735. ISSN: 0964-6906. DOI: 10.1093/hmg/ddg180. ppublish.

[21] V. V. Uslu, M. Petretich, S. Ruf, K. Langenfeld, N. A. Fonseca, J. C. Marioni and F. Spitz. Long-range enhancers regulating Myc expression are required for normal facial morphogenesis. *Nature genetics* 46 (7 July 2014), 753–758. ISSN: 1546-1718. DOI: 10.1038/ng.2971. ppublish.

[22] S. Schoenfelder, M. Furlan-Magaril, B. Mifsud, F. Tavares-Cadete, R. Sugar, B.-M. Javierre, T. Nagano, Y. Katsman, M. Sakthidevi, S. W. Wingett, E. Dimitrova, A. Dimond, L. B. Edelman, S. Elderkin, K. Tabbada, E. Darbo, S. Andrews, B. Herman, A. Higgs, E. LeProust, C. S. Osborne, J. A. Mitchell, N. M. Luscombe and P. Fraser. The pluripotent regulatory circuitry connecting pro-

moters to their long-range interacting elements. *Genome research* 25 (4 Apr. 2015), 582–597. ISSN: 1549-5469. DOI: 10.1101/gr.185272.114. ppublish.

[23] K. Lee, C. C.-S. Hsiung, P. Huang, A. Raj and G. A. Blobel. Dynamic enhancer-gene body contacts during transcription elongation. *Genes & Development* 29 (19 Oct. 2015), 1992–1997. ISSN: 1549-5477. DOI: 10.1101/gad.255265.114. ppublish.

[24] S. Schoenfelder and P. Fraser. Long-range enhancer-promoter contacts in gene expression control. *Nature reviews. Genetics* 20 (8 Aug. 2019), 437–455. ISSN: 1471-0064. DOI: 10.1038/s41576-019-0128-0. ppublish.

[25] G. Fudenberg, M. Imakaev, C. Lu, A. Goloborodko, N. Abdennur and L. A. Mirny. Formation of Chromosomal Domains by Loop Extrusion. *Cell reports* 15 (9 May 2016), 2038–2049. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2016.04.085. ppublish.

[26] Q. Szabo, F. Bantignies and G. Cavalli. Principles of genome folding into topologically associating domains. *Science advances* 5 (4 Apr. 2019), eaaw1668. ISSN: 2375-2548. DOI: 10.1126/sciadv.aaw1668. epublish.

[27] O. Symmons, V. V. Uslu, T. Tsujimura, S. Ruf, S. Nassari, W. Schwarzer, L. Ettwiller and F. Spitz. Functional and topological characteristics of mammalian regulatory domains. *Genome research* 24 (3 Mar. 2014), 390–400. ISSN: 1549-5469. DOI: 10.1101/gr.163519.113. ppublish.

[28] J. M. Dowen, Z. P. Fan, D. Hnisz, G. Ren, B. J. Abraham, L. N. Zhang, A. S. Weintraub, J. Schujiers, T. I. Lee, K. Zhao and R. A. Young. Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* 159 (2 Oct. 2014), 374–387. ISSN: 1097-4172. DOI: 10.1016/j.cell.2014.09.030. ppublish.

[29] D. Hnisz, A. S. Weintraub, D. S. Day, A.-L. Valton, R. O. Bak, C. H. Li, J. Goldmann, B. R. Lajoie, Z. P. Fan, A. A. Sigova, J. Reddy, D. Borges-Rivera, T. I. Lee, R. Jaenisch, M. H. Porteus, J. Dekker and R. A. Young. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science (New York, N.Y.)* 351 (6280 Mar. 2016), 1454–1458. ISSN: 1095-9203. DOI: 10.1126/science.aad9024. ppublish.

[30] X. Ji, D. B. Dadon, B. E. Powell, Z. P. Fan, D. Borges-Rivera, S. Shachar, A. S. Weintraub, D. Hnisz, G. Pegoraro, T. I. Lee, T. Misteli, R. Jaenisch and R. A. Young. 3D Chromosome Regulatory Landscape of Human Pluripotent Cells. *Cell stem cell* 18 (2 Feb. 2016), 262–275. ISSN: 1875-9777. DOI: 10.1016/j.stem.2015.11.007. ppublish.

[31] D. Hnisz, D. S. Day and R. A. Young. Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control. *Cell* 167 (5 Nov. 2016), 1188–1200. ISSN: 1097-4172. DOI: 10.1016/j.cell.2016.10.024. ppublish.

[32] J. Wang, J. Zhuang, S. Iyer, X. Lin, T. W. Whitfield, M. C. Greven, B. G. Pierce, X. Dong, A. Kundaje, Y. Cheng et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome research* 22.9 (2012), 1798–1812.

[33] T. Siggers and R. Gordaˆn. Protein–DNA binding: complexities and multi-protein codes. *Nucleic acids research* 42.4 (2014), 2099–2111.

[34] E. Morgunova and J. Taipale. Structural perspective of cooperative transcription factor binding. *Current opinion in structural biology* 47 (2017), 1–8.

[35] M. Levo and E. Segal. In pursuit of design principles of regulatory sequences. *Nature Reviews Genetics* 15.7 (2014), 453–468.

[36] R. L. Siegel, K. D. Miller and A. Jemal. Cancer statistics, 2020. *CA: a cancer journal for clinicians* 70 (1 Jan. 2020), 7–30. ISSN: 1542-4863. DOI: 10.3322/caac.21590. ppublish.

[37] A. L. Potosky, B. A. Miller, P. C. Albertsen and B. S. Kramer. The role of increasing detection in the rising incidence of prostate cancer. *JAMA* 273 (7 Feb. 1995), 548–552. ISSN: 0098-7484. ppublish.

[38] C. H. Pernar, E. M. Ebot, K. M. Wilson and L. A. Mucci. The Epidemiology of Prostate Cancer. *Cold Spring Harbor perspectives in medicine* 8 (12 Dec. 2018). ISSN: 2157-1422. DOI: 10.1101/cshperspect.a030361. epublish.

[39] J. C. Weinreb, J. O. Barentsz, P. L. Choyke, F. Cornud, M. A. Haider, K. J. Macura, D. Margolis, M. D. Schnall, F. Shtern, C. M. Tempany, H. C. Thoeny and S. Verma. PI-RADS Prostate Imaging - Reporting and Data System: 2015, Version 2. *European urology* 69 (1 Jan. 2016), 16–40. ISSN: 1873-7560. DOI: 10.1016/j.eururo.2015.08.052. ppublish.

[40] S. Woo, C. H. Suh, S. Y. Kim, J. Y. Cho and S. H. Kim. Diagnostic Performance of Prostate Imaging Reporting and Data System Version 2 for Detection of Prostate Cancer: A Systematic Review and Diagnostic Meta-analysis. *European urology* 72 (2 Aug. 2017), 177–188. ISSN: 1873-7560. DOI: 10.1016/j.eururo.2017.01.042. ppublish.

[41] M. M. Siddiqui, S. Rais-Bahrami, B. Turkbey, A. K. George, J. Rothwax, N. Shakir, C. Okoro, D. Raskolnikov, H. L. Parnes, W. M. Linehan, M. J. Merino, R. M. Simon, P. L. Choyke, B. J. Wood and P. A. Pinto. Comparison of MR/ultrasound fusion-guided biopsy with ultrasound-guided biopsy for the diagnosis of prostate cancer. *JAMA* 313 (4 Jan. 2015), 390–397. ISSN: 1538-3598. DOI: 10.1001/jama.2014.17942. ppublish.

[42] J. T. Wei, Z. Feng, A. W. Partin, E. Brown, I. Thompson, L. Sokoll, D. W. Chan, Y. Lotan, A. S. Kibel, J. E. Busby, M. Bidair, D. W. Lin, S. S. Taneja, R. Viterbo, A. Y. Joon, J. Dahlgren, J. Kagan, S. Srivastava and M. G. Sanda. Can urinary PCA3 supplement PSA in the early detection of prostate cancer?: *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 32 (36 Dec. 2014), 4066–4072. ISSN: 1527-7755. DOI: 10.1200/JCO.2013.52.8505. ppublish.

[43] A. W. Partin, L. Van Neste, E. A. Klein, L. S. Marks, J. R. Gee, D. A. Troyer, K. Rieger-Christ, J. S. Jones, C. Magi-Galluzzi, L. A. Mangold, B. J. Trock, R. S. Lance, J. W. Bigley, W. Van Criekinge and J. I. Epstein. Clinical validation of an epigenetic assay to predict negative histopathological results in repeat prostate biopsies. *The Journal of urology* 192 (4 Oct. 2014), 1081–1087. ISSN: 1527-3792. DOI: 10.1016/j.juro.2014.04.013. ppublish.

[44] P. Ström, K. Kartasalo, H. Olsson, L. Solorzano, B. Delahunt, D. M. Berney, D. G. Bostwick, A. J. Evans, D. J. Grignon, P. A. Humphrey, K. A. Iczkowski, J. G. Kench, G. Kristiansen, T. H. van der Kwast, K. R. M. Leite, J. K. McKenney, J. Oxley, C.-.-C. Pan, H. Samaratunga, J. R. Srigley, H. Takahashi, T. Tsuzuki, M. Varma, M. Zhou, J. Lindberg, C. Lindskog, P. Ruusuvuori, C. Wählby, H. Grönberg, M. Rantalainen, L. Egevad and M. Eklund. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *The Lancet. Oncology* 21 (2 Feb. 2020), 222–232. ISSN: 1474-5488. DOI: 10.1016/S1470-2045(19)30738-7. ppublish.

[45] S. Taavitsainen, M. Annala, E. Ledet, K. Beja, P. J. Miller, M. Moses, M. Nykter, K. N. Chi, O. Sartor and A. W. Wyatt. Evaluation of Commercial Circulating Tumor DNA Test in Metastatic Prostate Cancer. *JCO Precision Oncology* 3 (Nov. 2019), 1–9. DOI: 10.1200/po.19.00014.

[46] M. Annala, G. Vandekerkhove, D. Khalaf, S. Taavitsainen, K. Beja, E. W. Warner, K. Sunderland, C. Kollmannsberger, B. J. Eigl, D. Finch, C. D. Oja, J. Vergidis, M. Zulfiqar, A. A. Azad, M. Nykter, M. E. Gleave, A. W. Wyatt and K. N. Chi. Circulating Tumor DNA Genomics Correlate with Resistance to Abiraterone and Enzalutamide in Prostate Cancer. *Cancer discovery* 8 (4 Apr. 2018), 444–457. ISSN: 2159-8290. DOI: 10.1158/2159-8290.CD-17-0937. ppublish.

[47] G. Vandekerkhove, W. J. Struss, M. Annala, H. M. L. Kallio, D. Khalaf, E. W. Warner, C. Herberts, E. Ritch, K. Beja, Y. Loktionova, A. Hurtado-Coll, L. Fazli, A. So, P. C. Black, M. Nykter, T. Tammela, K. N. Chi, M. E. Gleave and A. W. Wyatt. Circulating Tumor DNA Abundance and Potential Utility in De Novo Metastatic Prostate Cancer. *European urology* 75 (4 Apr. 2019), 667–675. ISSN: 1873-7560. DOI: 10.1016/j.eururo.2018.12.042. ppublish.

[48] T. J. Daskivich, K.-H. Fan, T. Koyama, P. C. Albertsen, M. Goodman, A. S. Hamilton, R. M. Hoffman, J. L. Stanford, A. M. Stroup, M. S. Litwin and D. F. Penson. Effect of age, tumor risk, and comorbidity on competing risks for survival in a U.S. population-based cohort of men with prostate cancer. *Annals of internal medicine* 158 (10 May 2013), 709–717. ISSN: 1539-3704. DOI: 10.7326/0003-4819-158-10-201305210-00005. ppublish.

[49] M. S. Litwin and H.-J. Tan. The Diagnosis and Treatment of Prostate Cancer: A Review. *JAMA* 317 (24 June 2017), 2532–2542. ISSN: 1538-3598. DOI: 10.1001/jama.2017.7248. ppublish.

[50] G. S. Bova, H. M. L. Kallio, M. Annala, K. Kivinummi, G. Högnäs, S. Häyrynen, T. Rantapero, V. Kivinen, W. B. Isaacs, T. Tolonen, M. Nykter and T. Visakorpi. Integrated clinical, whole-genome, and transcriptome analysis of multisampled lethal metastatic prostate cancer. *Cold Spring Harbor molecular case studies* 2 (3 May 2016), a000752. ISSN: 2373-2873. DOI: 10.1101/mcs.a000752. ppublish.

[51] G. Gundem, P. Van Loo, B. Kremeyer, L. B. Alexandrov, J. M. C. Tubio, E. Papaemmanuil, D. S. Brewer, H. M. L. Kallio, G. Högnäs, M. Annala, K. Kivinummi, V. Goody, C. Latimer, S. O'Meara, K. J. Dawson, W. Isaacs, M. R. Emmert-Buck, M. Nykter, C. Foster, Z. Kote-Jarai, D. Easton, H. C. Whitaker, I. P. Group, D. E. Neal, C. S. Cooper, R. A. Eeles, T. Visakorpi, P. J. Campbell, U. McDermott, D. C. Wedge and G. S. Bova. The evolutionary history of lethal metastatic prostate cancer. *Nature* 520 (7547 Apr. 2015), 353–357. ISSN: 1476-4687. DOI: 10.1038/nature14347. ppublish.

[52] N. B. Atkin and M. C. Baker. Chromosome 10 deletion in carcinoma of the prostate. *The New England journal of medicine* 312 (5 Jan. 1985), 315. ISSN: 0028-4793. DOI: 10.1056/NEJM198501313120515. ppublish.

[53] J. J. König, E. Kamst, A. Hagemeijer, J. C. Romijn, J. Horoszewicz and F. H. Schröder. Cytogenetic characterization of several androgen responsive and unresponsive sublines of the human prostatic carcinoma cell line LNCaP. *Urological research* 17 (2 1989), 79–86. ISSN: 0300-5623. DOI: 10.1007/BF00262025. ppublish.

[54] S. J. Rubin, D. E. Hallahan, C. R. Ashman, D. G. Brachman, M. A. Beckett, S. Virudachalam, D. W. Yandell and R. R. Weichselbaum. Two prostate carcinoma cell lines demonstrate abnormalities in tumor suppressor genes. *Journal of surgical oncology* 46 (1 Jan. 1991), 31–36. ISSN: 0022-4790. DOI: 10.1002/jso.2930460108. ppublish.

[55] R. Lundgren, N. Mandahl, S. Heim, J. Limon, H. Henrikson and F. Mitelman. Cytogenetic analysis of 57 primary prostatic adenocarcinomas. *Genes, chromosomes & cancer* 4 (1 Jan. 1992), 16–24. ISSN: 1045-2257. DOI: 10.1002/gcc.2870040103. ppublish.

[56] B. S. Carter, C. M. Ewing, W. S. Ward, B. F. Treiger, T. W. Aalders, J. A. Schalken, J. I. Epstein and W. B. Isaacs. Allelic loss of chromosomes 16q and 10q in human prostate cancer. *Proceedings of the National Academy of Sciences of the United States of America* 87 (22 Nov. 1990), 8751–8755. ISSN: 0027-8424. DOI: 10.1073/pnas.87.22.8751. ppublish.

[57] R. Bookstein, J. Y. Shew, P. L. Chen, P. Scully and W. H. Lee. Suppression of tumorigenicity of human prostate carcinoma cells by replacing a mutated RB

gene. *Science (New York, N.Y.)* 247 (4943 Feb. 1990), 712–715. ISSN: 0036-8075. DOI: 10.1126/science.2300823. ppublish.

[58] G. S. Bova, D. MacGrogan, A. Levy, S. S. Pin, R. Bookstein and W. B. Isaacs. Physical mapping of chromosome 8p22 markers and their homozygous deletion in a metastatic prostate cancer. *Genomics* 35 (1 July 1996), 46–54. ISSN: 0888-7543. DOI: 10.1006/geno.1996.0321. ppublish.

[59] W. W. He, P. J. Sciavolino, J. Wing, M. Augustus, P. Hudson, P. S. Meissner, R. T. Curtis, B. K. Shell, D. G. Bostwick, D. J. Tindall, E. P. Gelmann, C. Abate-Shen and K. C. Carter. A novel human prostate-specific, androgen-regulated homeobox gene (NKX3.1) that maps to 8p21, a region frequently deleted in prostate cancer. *Genomics* 43 (1 July 1997), 69–77. ISSN: 0888-7543. DOI: 10.1006/geno.1997.4715. ppublish.

[60] J. R. Newmark, D. O. Hardy, D. C. Tonb, B. S. Carter, J. I. Epstein, W. B. Isaacs, T. R. Brown and E. R. Barrack. Androgen receptor gene mutations in human prostate cancer. *Proceedings of the National Academy of Sciences of the United States of America* 89 (14 July 1992), 6319–6323. ISSN: 0027-8424. DOI: 10.1073/pnas.89.14.6319. ppublish.

[61] M. E. Taplin, G. J. Bubley, T. D. Shuster, M. E. Frantz, A. E. Spooner, G. K. Ogata, H. N. Keer and S. P. Balk. Mutation of the androgen-receptor gene in metastatic androgen-independent prostate cancer. *The New England journal of medicine* 332 (21 May 1995), 1393–1398. ISSN: 0028-4793. DOI: 10.1056/NEJM199505253322101. ppublish.

[62] J. Li, C. Yen, D. Liaw, K. Podsypanina, S. Bose, S. I. Wang, J. Puc, C. Miliaresis, L. Rodgers, R. McCombie, S. H. Bigner, B. C. Giovanella, M. Ittmann, B. Tycko, H. Hibshoosh, M. H. Wigler and R. Parsons. PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. *Science (New York, N.Y.)* 275 (5308 Mar. 1997), 1943–1947. ISSN: 0036-8075. DOI: 10.1126/science.275.5308.1943. ppublish.

[63] W. H. Fleming, A. Hamel, R. MacDonald, E. Ramsey, N. M. Pettigrew, B. Johnston, J. G. Dodd and R. J. Matusik. Expression of the c-myc protooncogene in human prostatic carcinoma and benign prostatic hyperplasia. *Cancer research* 46 (3 Mar. 1986), 1535–1538. ISSN: 0008-5472. ppublish.

[64] K. Ellwood-Yen, T. G. Graeber, J. Wongvipat, M. L. Iruela-Arispe, J. Zhang, R. Matusik, G. V. Thomas and C. L. Sawyers. Myc-driven murine prostate cancer shares molecular features with human prostate tumors. *Cancer cell* 4 (3 Sept. 2003), 223–238. ISSN: 1535-6108. DOI: 10.1016/s1535-6108(03)00197-1. ppublish.

[65] N. J. Clegg, S. S. Couto, J. Wongvipat, H. Hieronymus, B. S. Carver, B. S. Taylor, K. Ellwood-Yen, W. L. Gerald, C. Sander and C. L. Sawyers. MYC cooperates with AKT in prostate tumorigenesis and alters sensitivity to mTOR inhibitors. *PloS one* 6 (3 Mar. 2011), e17449. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0017449. epublish.

[66] T. Visakorpi, A. H. Kallioniemi, A. C. Syvänen, E. R. Hyytinen, R. Karhu, T. Tammela, J. J. Isola and O. P. Kallioniemi. Genetic changes in primary and recurrent prostate cancer by comparative genomic hybridization. *Cancer research* 55 (2 Jan. 1995), 342–347. ISSN: 0008-5472. ppublish.

[67] M. L. Cher, D. MacGrogan, R. Bookstein, J. A. Brown, R. B. Jenkins and R. H. Jensen. Comparative genomic hybridization, allelic imbalance, and fluorescence in situ hybridization on chromosome 8 in prostate cancer. *Genes, chromosomes & cancer* 11 (3 Nov. 1994), 153–162. ISSN: 1045-2257. DOI: 10.1002/gcc.2870110304. ppublish.

[68] S. A. Tomlins, D. R. Rhodes, S. Perner, S. M. Dhanasekaran, R. Mehra, X.-W. Sun, S. Varambally, X. Cao, J. Tchinda, R. Kuefer, C. Lee, J. E. Montie, R. B. Shah, K. J. Pienta, M. A. Rubin and A. M. Chinnaiyan. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science (New York, N.Y.)* 310 (5748 Oct. 2005), 644–648. ISSN: 1095-9203. DOI: 10.1126/science.1117679. ppublish.

[69] S. A. Tomlins, R. Mehra, D. R. Rhodes, L. R. Smith, D. Roulston, B. E. Helgeson, X. Cao, J. T. Wei, M. A. Rubin, R. B. Shah and A. M. Chinnaiyan. TMPRSS2:ETV4 gene fusions define a third molecular subtype of prostate cancer. *Cancer research* 66 (7 Apr. 2006), 3396–3400. ISSN: 0008-5472. DOI: 10.1158/0008-5472.CAN-06-0168. ppublish.

[70] P. Paulo, J. D. Barros-Silva, F. R. Ribeiro, J. Ramalho-Carvalho, C. Jerónimo, R. Henrique, G. E. Lind, R. I. Skotheim, R. A. Lothe and M. R. Teixeira. FLI1 is a novel ETS transcription factor involved in gene fusions in prostate cancer.

Genes, chromosomes & cancer 51 (3 Mar. 2012), 240–249. ISSN: 1098-2264. DOI: 10.1002/gcc.20948. ppublish.

[71] M. A. Rubin and F. Demichelis. The Genomics of Prostate Cancer: A Historic Perspective. *Cold Spring Harbor perspectives in medicine* 9 (3 Mar. 2019). ISSN: 2157-1422. DOI: 10.1101/cshperspect.a034942. epublish.

[72] C. G. A. R. Network. The Molecular Taxonomy of Primary Prostate Cancer. *Cell* 163 (4 Nov. 2015), 1011–1025. ISSN: 1097-4172. DOI: 10.1016/j.cell.2015.10.025. ppublish.

[73] C. Abate-Shen and M. M. Shen. Molecular genetics of prostate cancer. *Genes & development* 14 (19 Oct. 2000), 2410–2434. ISSN: 0890-9369. DOI: 10.1101/gad.819500. ppublish.

[74] S.-Y. Ku, M. E. Gleave and H. Beltran. Towards precision oncology in advanced prostate cancer. *Nature reviews. Urology* 16 (11 Nov. 2019), 645–654. ISSN: 1759-4820. DOI: 10.1038/s41585-019-0237-8. ppublish.

[75] J. Trapman, C. Ris-Stalpers, J. A. van der Korput, G. G. Kuiper, P. W. Faber, J. C. Romijn, E. Mulder and A. O. Brinkmann. The androgen receptor: functional structure and expression in transplanted human prostate tumors and prostate tumor cell lines. *The Journal of steroid biochemistry and molecular biology* 37 (6 Dec. 1990), 837–842. ISSN: 0960-0760. DOI: 10.1016/0960-0760(90)90429-o. ppublish.

[76] J. P. Gaddipati, D. G. McLeod, H. B. Heidenberg, I. A. Sesterhenn, M. J. Finger, J. W. Moul and S. Srivastava. Frequent detection of codon 877 mutation in the androgen receptor gene in advanced prostate cancers. *Cancer research* 54 (11 June 1994), 2861–2864. ISSN: 0008-5472. ppublish.

[77] J. P. Elo, L. Kvist, K. Leinonen, V. Isomaa, P. Henttu, O. Lukkarinen and P. Vihko. Mutated human androgen receptor gene detected in a prostatic cancer patient is also activated by estradiol. *The Journal of clinical endocrinology and metabolism* 80 (12 Dec. 1995), 3494–3500. ISSN: 0021-972X. DOI: 10.1210/jcem.80.12.8530589. ppublish.

[78] J. Veldscholte, C. A. Berrevoets, C. Ris-Stalpers, G. G. Kuiper, G. Jenster, J. Trapman, A. O. Brinkmann and E. Mulder. The androgen receptor in LNCaP cells contains a mutation in the ligand binding domain which affects steroid

binding characteristics and response to antiandrogens. *The Journal of steroid biochemistry and molecular biology* 41 (3-8 Mar. 1992), 665–669. ISSN: 0960-0760. DOI: 10.1016/0960-0760(92)90401-4. ppublish.

[79] W. D. Tilley, G. Buchanan, T. E. Hickey and J. M. Bentel. Mutations in the androgen receptor gene are associated with progression of human prostate cancer to androgen independence. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2 (2 Feb. 1996), 277–285. ISSN: 1078-0432. ppublish.

[80] J. A. Ruizeveld de Winter, P. J. Janssen, H. M. Sleddens, M. C. Verleun-Mooijman, J. Trapman, A. O. Brinkmann, A. B. Santerse, F. H. Schröder and T. H. van der Kwast. Androgen receptor status in localized and locally progressive hormone refractory human prostate cancer. *The American journal of pathology* 144 (4 Apr. 1994), 735–746. ISSN: 0002-9440. ppublish.

[81] S. R. Viswanathan, G. Ha, A. M. Hoff, J. A. Wala, J. Carrot-Zhang, C. W. Whelan, N. J. Haradhvala, S. S. Freeman, S. C. Reed, J. Rhoades, P. Polak, M. Cipicchio, S. A. Wankowicz, A. Wong, T. Kamath, Z. Zhang, G. J. Gydush, D. Rotem, P. I. P. C. D. Team, J. C. Love, G. Getz, S. Gabriel, C.-Z. Zhang, S. M. Dehm, P. S. Nelson, E. M. Van Allen, A. D. Choudhury, V. A. Adalsteinsson, R. Beroukhim, M.-E. Taplin and M. Meyerson. Structural Alterations Driving Castration-Resistant Prostate Cancer Revealed by Linked-Read Genome Sequencing. *Cell* 174 (2 July 2018), 433–447.e19. ISSN: 1097-4172. DOI: 10.1016/j.cell.2018.05.036. ppublish.

[82] D. A. Quigley, H. X. Dang, S. G. Zhao, P. Lloyd, R. Aggarwal, J. J. Alumkal, A. Foye, V. Kothari, M. D. Perry, A. M. Bailey, D. Playdle, T. J. Barnard, L. Zhang, J. Zhang, J. F. Youngren, M. P. Cieslik, A. Parolia, T. M. Beer, G. Thomas, K. N. Chi, M. Gleave, N. A. Lack, A. Zoubeidi, R. E. Reiter, M. B. Rettig, O. Witte, C. J. Ryan, L. Fong, W. Kim, T. Friedlander, J. Chou, H. Li, R. Das, H. Li, R. Moussavi-Baygi, H. Goodarzi, L. A. Gilbert, P. N. Lara, C. P. Evans, T. C. Goldstein, J. M. Stuart, S. A. Tomlins, D. E. Spratt, R. K. Cheetham, D. T. Cheng, K. Farh, J. S. Gehring, J. Hakenberg, A. Liao, P. G. Febbo, J. Shon, B. Sickler, S. Batzoglou, K. E. Knudsen, H. H. He, J. Huang, A. W. Wyatt, S. M. Dehm, A. Ashworth, A. M. Chinnaiyan, C. A. Maher, E. J. Small and F. Y. Feng. Genomic Hallmarks and Structural Variation in

Metastatic Prostate Cancer. *Cell* 174 (3 July 2018), 758–769.e9. ISSN: 1097-4172. DOI: 10.1016/j.cell.2018.06.039. ppublish.

[83] P. A. Watson, V. K. Arora and C. L. Sawyers. Emerging mechanisms of resistance to androgen receptor inhibitors in prostate cancer. *Nature reviews. Cancer* 15 (12 Dec. 2015), 701–711. ISSN: 1474-1768. DOI: 10.1038/nrc4016. ppublish.

[84] K.-H. Chang, R. Li, B. Kuri, Y. Lotan, C. G. Roehrborn, J. Liu, R. Vessella, P. S. Nelson, P. Kapur, X. Guo, H. Mirzaei, R. J. Auchus and N. Sharifi. A gain-of-function mutation in DHT synthesis in castration-resistant prostate cancer. *Cell* 154 (5 Aug. 2013), 1074–1084. ISSN: 1097-4172. DOI: 10.1016/j.cell.2013.07.029. ppublish.

[85] D. Robinson, E. M. Van Allen, Y.-M. Wu, N. Schultz, R. J. Lonigro, J.-M. Mosquera, B. Montgomery, M.-E. Taplin, C. C. Pritchard, G. Attard, H. Beltran, W. Abida, R. K. Bradley, J. Vinson, X. Cao, P. Vats, L. P. Kunju, M. Hussain, F. Y. Feng, S. A. Tomlins, K. A. Cooney, D. C. Smith, C. Brennan, J. Siddiqui, R. Mehra, Y. Chen, D. E. Rathkopf, M. J. Morris, S. B. Solomon, J. C. Durack, V. E. Reuter, A. Gopalan, J. Gao, M. Loda, R. T. Lis, M. Bowden, S. P. Balk, G. Gaviola, C. Sougnez, M. Gupta, E. Y. Yu, E. A. Mostaghel, H. H. Cheng, H. Mulcahy, L. D. True, S. R. Plymate, H. Dvinge, R. Ferraldeschi, P. Flohr, S. Miranda, Z. Zafeiriou, N. Tunariu, J. Mateo, R. Perez-Lopez, F. Demichelis, B. D. Robinson, M. Schiffman, D. M. Nanus, S. T. Tagawa, A. Sigaras, K. W. Eng, O. Elemento, A. Sboner, E. I. Heath, H. I. Scher, K. J. Pienta, P. Kantoff, J. S. de Bono, M. A. Rubin, P. S. Nelson, L. A. Garraway, C. L. Sawyers and A. M. Chinnaiyan. Integrative clinical genomics of advanced prostate cancer. *Cell* 161 (5 May 2015), 1215–1228. ISSN: 1097-4172. DOI: 10.1016/j.cell.2015.05.001. ppublish.

[86] R. Bookstein, D. MacGrogan, S. G. Hilsenbeck, F. Sharkey and D. C. Allred. p53 is mutated in a subset of advanced-stage prostate cancers. *Cancer research* 53.14 (1993), 3369–3373.

[87] P. J. Effert, A. Neubauer, P. J. Walther and E. T. Liu. Alterations of the P53 gene are associated with the progression of a human prostate carcinoma. *The Journal of urology* 147 (3 Pt 2 Mar. 1992), 789–793. ISSN: 0022-5347. DOI: 10.1016/s0022-5347(17)37387-1. ppublish.

[88] N. M. Navone, P. Troncoso, L. L. Pisters, T. L. Goodrow, J. L. Palmer, W. W. Nichols, A. C. von Eschenbach and C. J. Conti. p53 protein accumulation and gene mutation in the progression of human prostate carcinoma. *Journal of the National Cancer Institute* 85 (20 Oct. 1993), 1657–1669. ISSN: 0027-8874. DOI: 10.1093/jnci/85.20.1657. ppublish.

[89] A. G. Aprikian, A. S. Sarkis, W. R. Fair, Z. F. Zhang, Z. Fuks and C. Cordon-Cardo. Immunohistochemical determination of p53 protein nuclear accumulation in prostatic adenocarcinoma. *The Journal of urology* 151 (5 May 1994), 1276–1280. ISSN: 0022-5347. DOI: 10.1016/s0022-5347(17)35231-x. ppublish.

[90] J. A. Eastham, A. M. Stapleton, A. E. Gousse, T. L. Timme, G. Yang, K. M. Slawin, T. M. Wheeler, P. T. Scardino and T. C. Thompson. Association of p53 mutations with metastatic prostate cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research* 1 (10 Oct. 1995), 1111–1118. ISSN: 1078-0432. ppublish.

[91] H. B. Heidenberg, I. A. Sesterhenn, J. P. Gaddipati, C. M. Weghorst, G. S. Buzard, J. W. Moul and S. Srivastava. Alteration of the tumor suppressor gene p53 in a high fraction of hormone refractory prostate cancer. *The Journal of urology* 154 (2 Pt 1 Aug. 1995), 414–421. ISSN: 0022-5347. DOI: 10.1097/00005392-199508000-00024. ppublish.

[92] M. Colombel, F. Symmans, S. Gil, K. M. O'Toole, D. Chopin, M. Benson, C. A. Olsson, S. Korsmeyer and R. Buttyan. Detection of the apoptosis-suppressing oncoprotein bc1-2 in hormone-refractory human prostate cancers. *The American journal of pathology* 143 (2 Aug. 1993), 390–400. ISSN: 0002-9440. ppublish.

[93] I. Apakama, M. C. Robinson, N. M. Walter, R. G. Charlton, J. A. Royds, C. E. Fuller, D. E. Neal and F. C. Hamdy. bcl-2 overexpression combined with p53 protein accumulation correlates with hormone-refractory prostate cancer. *British journal of cancer* 74 (8 Oct. 1996), 1258–1262. ISSN: 0007-0920. DOI: 10.1038/bjc.1996.526. ppublish.

[94] Y. Furuya, S. Krajewski, J. I. Epstein, J. C. Reed and J. T. Isaacs. Expression of bcl-2 and the progression of human and rodent prostatic cancers. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2 (2 Feb. 1996), 389–398. ISSN: 1078-0432. ppublish.

[95] T. J. McDonnell, N. M. Navone, P. Troncoso, L. L. Pisters, C. Conti, A. C. von Eschenbach, S. Brisbay and C. J. Logothetis. Expression of bcl-2 oncoprotein and p53 protein accumulation in bone marrow metastases of androgen independent prostate cancer. *The Journal of urology* 157 (2 Feb. 1997), 569–574. ISSN: 0022-5347. ppublish.

[96] S. M. Tu, K. McConnell, M. C. Marin, M. L. Campbell, A. Fernandez, A. C. von Eschenbach and T. J. McDonnell. Combination adriamycin and suramin induces apoptosis in bcl-2 expressing prostate carcinoma cells. *Cancer letters* 93 (2 July 1995), 147–155. ISSN: 0304-3835. DOI: 10.1016/0304-3835(95)03795-x. ppublish.

[97] H. J. Voeller, C. I. Truica and E. P. Gelmann. Beta-catenin mutations in human prostate cancer. *Cancer research* 58 (12 June 1998), 2520–2523. ISSN: 0008-5472. ppublish.

[98] L. F. van Dessel, J. van Riet, M. Smits, Y. Zhu, P. Hamberg, M. S. van der Heijden, A. M. Bergman, I. M. van Oort, R. de Wit, E. E. Voest, N. Steeghs, T. N. Yamaguchi, J. Livingstone, P. C. Boutros, J. W. M. Martens, S. Sleijfer, E. Cuppen, W. Zwart, H. J. G. van de Werken, N. Mehra and M. P. Lolkema. The genomic landscape of metastatic castration-resistant prostate cancers reveals multiple distinct genotypes with potential clinical impact. *Nature communications* 10 (1 Nov. 2019), 5251. ISSN: 2041-1723. DOI: 10.1038/s41467-019-13084-7. epublish.

[99] C. B. Anfinsen, E. Haber, S. E. L. A. M and F. H. White. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences of the United States of America* 47 (Sept. 1961), 1309–1314. ISSN: 0027-8424. DOI: 10.1073/pnas.47.9.1309. ppublish.

[100] P. E. Wright and H. J. Dyson. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of molecular biology* 293 (2 Oct. 1999), 321–331. ISSN: 0022-2836. DOI: 10.1006/jmbi.1999.3110. ppublish.

[101] P. Tompa. Intrinsically unstructured proteins. *Trends in biochemical sciences* 27.10 (2002), 527–533.

[102] V. N. Uversky, J. R. Gillespie and A. L. Fink. Why are "natively unfolded" proteins unstructured under physiologic conditions?: *Proteins* 41 (3 Nov. 2000), 415–427. ISSN: 0887-3585. DOI: 10.1002/1097-0134(20001115)41:3<415::aid-prot130>3.0.co;2-7. ppublish.

[103] V. N. Uversky. Natively unfolded proteins: a point where biology waits for physics. *Protein science : a publication of the Protein Society* 11 (4 Apr. 2002), 739–756. ISSN: 0961-8368. DOI: 10.1110/ps.4210102. ppublish.

[104] A. L. Fink. Natively unfolded proteins. *Current opinion in structural biology* 15 (1 Feb. 2005), 35–41. ISSN: 0959-440X. DOI: 10.1016/j.sbi.2005.01.002. ppublish.

[105] H. J. Dyson and P. E. Wright. Intrinsically unstructured proteins and their functions. *Nature reviews. Molecular cell biology* 6 (3 Mar. 2005), 197–208. ISSN: 1471-0072. DOI: 10.1038/nrm1589. ppublish.

[106] R. Kumar and E. B. Thompson. Transactivation functions of the N-terminal domains of nuclear hormone receptors: protein folding and coactivator interactions. *Molecular endocrinology (Baltimore, Md.)* 17 (1 Jan. 2003), 1–10. ISSN: 0888-8809. DOI: 10.1210/me.2002-0258. ppublish.

[107] A. K. Dunker, J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff, K. W. Hipps, J. Ausio, M. S. Nissen, R. Reeves, C. Kang, C. R. Kissinger, R. W. Bailey, M. D. Griswold, W. Chiu, E. C. Garner and Z. Obradovic. Intrinsically disordered protein. *Journal of molecular graphics & modelling* 19 (1 2001), 26–59. ISSN: 1093-3263. DOI: 10.1016/s1093-3263(00)00138-8. ppublish.

[108] A. K. Dunker and Z. Obradovic. The protein trinity–linking function and disorder. *Nature biotechnology* 19 (9 Sept. 2001), 805–806. ISSN: 1087-0156. DOI: 10.1038/nbt0901-805. ppublish.

[109] S. L. Flaugh and K. J. Lumb. Effects of macromolecular crowding on the intrinsically disordered proteins c-Fos and p27(Kip1). *Biomacromolecules* 2 (2 **Summer** 2001), 538–540. ISSN: 1525-7797. DOI: 10.1021/bm015502z. ppublish.

[110] K. Namba. Roles of partly unfolded conformations in macromolecular self-assembly. *Genes to cells : devoted to molecular & cellular mechanisms* 6 (1 Jan. 2001), 1–12. ISSN: 1356-9597. DOI: 10.1046/j.1365-2443.2001.00384.x. ppublish.

[111] V. N. Uversky. A protein-chameleon: conformational plasticity of alpha-synuclein, a disordered protein involved in neurodegenerative disorders. *Journal of biomolecular structure & dynamics* 21 (2 Oct. 2003), 211–234. ISSN: 0739-1102. DOI: 10.1080/07391102.2003.10506918. ppublish.

[112] V. N. Uversky, C. J. Oldfield and A. K. Dunker. Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *Journal of molecular recognition : JMR* 18 (5 2005), 343–384. ISSN: 0952-3499. DOI: 10.1002/jmr.747. ppublish.

[113] H. Xie, S. Vucetic, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, Z. Obradovic and V. N. Uversky. Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins. *Journal of proteome research* 6 (5 May 2007), 1917–1932. ISSN: 1535-3893. DOI: 10.1021/pr060394e. ppublish.

[114] H. Xie, S. Vucetic, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, V. N. Uversky and Z. Obradovic. Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *Journal of proteome research* 6 (5 May 2007), 1882–1898. ISSN: 1535-3893. DOI: 10.1021/pr060392u. ppublish.

[115] S. Vucetic, H. Xie, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, Z. Obradovic and V. N. Uversky. Functional anthology of intrinsic disorder. 2. Cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions. *Journal of proteome research* 6 (5 May 2007), 1899–1916. ISSN: 1535-3893. DOI: 10.1021/pr060393m. ppublish.

[116] P. Tompa. The interplay between structure and function in intrinsically unstructured proteins. *FEBS letters* 579 (15 June 2005), 3346–3354. ISSN: 0014-5793. DOI: 10.1016/j.febslet.2005.03.072. ppublish.

[117] J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton and D. T. Jones. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *Journal of molecular biology* 337 (3 Mar. 2004), 635–645. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2004.02.002. ppublish.

[118] A. K. Dunker, E. Garner, S. Guilliot, P. Romero, K. Albrecht, J. Hart, Z. Obradovic, C. Kissinger and J. E. Villafranca. Protein disorder and the evolution of molecular recognition: theory, predictions and observations. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (1998), 473–484. ISSN: 2335-6928. ppublish.

[119] A. K. Dunker, Z. Obradovic, P. Romero, E. C. Garner and C. J. Brown. Intrinsic protein disorder in complete genomes. *Genome informatics. Workshop on Genome Informatics* 11 (2000), 161–171. ppublish.

[120] P. Romero, Z. Obradovic, C. R. Kissinger, J. E. Villafranca, E. Garner, S. Guilliot and A. K. Dunker. Thousands of proteins likely to have long disordered regions. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (1998), 437–448. ISSN: 2335-6928. ppublish.

[121] J. Vymětal, J. Vondrášek and K. Hlouchová. Sequence Versus Composition: What Prescribes IDP Biophysical Properties?: *Entropy* 21.7 (2019), 654.

[122] L. M. Iakoucheva, C. J. Brown, J. D. Lawson, Z. Obradović and A. K. Dunker. Intrinsic disorder in cell-signaling and cancer-associated proteins. *Journal of molecular biology* 323 (3 Oct. 2002), 573–584. ISSN: 0022-2836. DOI: 10.1016/s0022-2836(02)00969-5. ppublish.

[123] J. Liu, N. B. Perumal, C. J. Oldfield, E. W. Su, V. N. Uversky and A. K. Dunker. Intrinsic disorder in transcription factors. *Biochemistry* 45 (22 June 2006), 6873–6888. ISSN: 0006-2960. DOI: 10.1021/bi0602718. ppublish.

[124] L. Staby, C. O'Shea, M. Willemoës, F. Theisen, B. B. Kragelund and K. Skriver. Eukaryotic transcription factors: paradigms of protein intrinsic disorder. *The Biochemical journal* 474 (15 July 2017), 2509–2532. ISSN: 1470-8728. DOI: 10.1042/BCJ20160631. epublish.

[125] A. Enthart, C. Klein, A. Dehner, M. Coles, G. Gemmecker, H. Kessler and F. Hagn. Solution structure and binding specificity of the p63 DNA binding domain. *Scientific reports* 6 (May 2016), 26707. ISSN: 2045-2322. DOI: 10.1038/srep26707. epublish.

[126] X. Robert and P. Gouet. Deciphering key features in protein structures with the new ENDscript server. *Nucleic acids research* 42 (Web Server issue July 2014), W320–W324. ISSN: 1362-4962. DOI: 10.1093/nar/gku316. ppublish.

[127] P. Kulkarni and V. N. Uversky. Intrinsically Disordered Proteins: The Dark Horse of the Dark Proteome. *Proteomics* 18 (21-22 Nov. 2018), e1800061. ISSN: 1615-9861. DOI: 10.1002/pmic.201800061. ppublish.

[128] N. Perdigão, J. Heinrich, C. Stolte, K. S. Sabir, M. J. Buckley, B. Tabor, B. Signal, B. S. Gloss, C. J. Hammang, B. Rost, A. Schafferhans and S. I. O'Donoghue. Unexpected features of the dark proteome. *Proceedings of the National Academy of Sciences of the United States of America* 112 (52 Dec. 2015), 15898–15903. ISSN: 1091-6490. DOI: 10.1073/pnas.1508380112. ppublish.

[129] A. K. Dunker, M. S. Cortese, P. Romero, L. M. Iakoucheva and V. N. Uversky. Flexible nets. The roles of intrinsic disorder in protein interaction networks. *The FEBS journal* 272 (20 Oct. 2005), 5129–5148. ISSN: 1742-464X. DOI: 10.1111/j.1742-4658.2005.04948.x. ppublish.

[130] Romero, Obradovic and Dunker. Sequence Data Analysis for Long Disordered Regions Prediction in the Calcineurin Family. *Genome informatics. Workshop on Genome Informatics* 8 (1997), 110–124. ppublish.

[131] P. Romero, Z. Obradovic, C. Kissinger, J. Villafranca and A. K. Dunker. Identifying disordered regions in proteins from amino acid sequence. *Proceedings of International Conference on Neural Networks (ICNN'97)*. Vol. 1. IEEE. 1997, 90–95.

[132] R. M. Williams, Z. Obradovi, V. Mathura, W. Braun, E. C. Garner, J. Young, S. Takayama, C. J. Brown and A. K. Dunker. The protein non-folding problem: amino acid determinants of intrinsic order and disorder. *Pacific Symposium on Biocomputing* (2001), 89–100. ISSN: 2335-6928. ppublish.

[133] V. N. Uversky. Paradoxes and wonders of intrinsic disorder: Stability of instability. *Intrinsically disordered proteins* 5 (1 2017), e1327757. ISSN: 2169-0707. DOI: 10.1080/21690707.2017.1327757. epublish.

[134] H. J. Dyson and P. E. Wright. Coupling of folding and binding for unstructured proteins. *Current opinion in structural biology* 12 (1 Feb. 2002), 54–60. ISSN: 0959-440X. DOI: 10.1016/s0959-440x(02)00289-0. ppublish.

[135] A. K. Dunker and V. N. Uversky. Signal transduction via unstructured protein conduits. *Nature chemical biology* 4 (4 Apr. 2008), 229–230. ISSN: 1552-4469. DOI: 10.1038/nchembio0408-229. ppublish.

[136] R. van der Lee, M. Buljan, B. Lang, R. J. Weatheritt, G. W. Daughdrill, A. K. Dunker, M. Fuxreiter, J. Gough, J. Gsponer, D. T. Jones, P. M. Kim, R. W. Kriwacki, C. J. Oldfield, R. V. Pappu, P. Tompa, V. N. Uversky, P. E. Wright and M. M. Babu. Classification of intrinsically disordered regions and proteins. *Chemical reviews* 114 (13 July 2014), 6589–6631. ISSN: 1520-6890. DOI: 10.1021/cr400525m. ppublish.

[137] L. M. Iakoucheva, P. Radivojac, C. J. Brown, T. R. O'Connor, J. G. Sikes, Z. Obradovic and A. K. Dunker. The importance of intrinsic disorder for protein phosphorylation. *Nucleic acids research* 32 (3 2004), 1037–1049. ISSN: 1362-4962. DOI: 10.1093/nar/gkh253. epublish.

[138] A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva and Z. Obradović. Intrinsic disorder and protein function. *Biochemistry* 41 (21 May 2002), 6573–6582. ISSN: 0006-2960. DOI: 10.1021/bi012159+. ppublish.

[139] R. Kumar and E. B. Thompson. The structure of the nuclear hormone receptors. *Steroids* 64 (5 May 1999), 310–319. ISSN: 0039-128X. DOI: 10.1016/s0039-128x(99)00014-8. ppublish.

[140] K. Sugase, H. J. Dyson and P. E. Wright. Mechanism of coupled folding and binding of an intrinsically disordered protein. *Nature* 447 (7147 June 2007), 1021–1025. ISSN: 1476-4687. DOI: 10.1038/nature05858. ppublish.

[141] P. Robustelli, S. Piana and D. E. Shaw. Mechanism of Coupled Folding-upon-Binding of an Intrinsically Disordered Protein. *Journal of the American Chemical Society* 142 (25 June 2020), 11092–11101. ISSN: 1520-5126. DOI: 10.1021/jacs.0c03217. ppublish.

[142] B. Schmidtgall, O. Chaloin, V. Bauer, M. Sumyk, C. Birck and V. Torbeev. Dissecting mechanism of coupled folding and binding of an intrinsically disordered protein by chemical synthesis of conformationally constrained analogues. *Chemical communications (Cambridge, England)* 53 (53 June 2017), 7369–7372. ISSN: 1364-548X. DOI: 10.1039/c7cc02276j. ppublish.

[143] R. W. Kriwacki, L. Hengst, L. Tennant, S. I. Reed and P. E. Wright. Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity. *Proceedings of the National Academy of Sciences of the United States of America* 93 (21 Oct. 1996), 11504–11509. ISSN: 0027-8424. DOI: 10.1073/pnas.93.21.11504. ppublish.

[144] H. Kitano. Biological robustness. *Nature reviews. Genetics* 5 (11 Nov. 2004), 826–837. ISSN: 1471-0056. DOI: 10.1038/nrg1471. ppublish.

[145] Y. J. K. Edwards, A. E. Lobley, M. M. Pentony and D. T. Jones. Insights into the regulation of intrinsically disordered proteins in the human proteome by analyzing sequence and gene expression data. *Genome biology* 10 (5 2009), R50. ISSN: 1474-760X. DOI: 10.1186/gb-2009-10-5-r50. ppublish.

[146] T. Vavouri, J. I. Semple, R. Garcia-Verdugo and B. Lehner. Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. *Cell* 138 (1 July 2009), 198–208. ISSN: 1097-4172. DOI: 10.1016/j.cell.2009.04.029. ppublish.

[147] M. M. Babu, R. van der Lee, N. S. de Groot and J. Gsponer. Intrinsically disordered proteins: regulation and disease. *Current opinion in structural biology* 21 (3 June 2011), 432–440. ISSN: 1879-033X. DOI: 10.1016/j.sbi.2011.03.011. ppublish.

[148] A. S. Krois, H. J. Dyson and P. E. Wright. Long-range regulation of p53 DNA binding by its intrinsically disordered N-terminal transactivation domain. *Proceedings of the National Academy of Sciences of the United States of America* 115 (48 Nov. 2018), E11302–E11310. ISSN: 1091-6490. DOI: 10.1073/pnas.1814051115. ppublish.

[149] M. J. Friedman, A. G. Shah, Z.-H. Fang, E. G. Ward, S. T. Warren, S. Li and X.-J. Li. Polyglutamine domain modulates the TBP-TFIIB interaction: implications for its normal function and neurodegeneration. *Nature neuroscience* 10 (12 Dec. 2007), 1519–1528. ISSN: 1097-6256. DOI: 10.1038/nn2011. ppublish.

[150] S. Chong, C. Dugast-Darzacq, Z. Liu, P. Dong, G. M. Dailey, C. Cattoglio, A. Heckert, S. Banala, L. Lavis, X. Darzacq and R. Tjian. Imaging dynamic and selective low-complexity domain interactions that control gene transcription. *Science (New York, N.Y.)* 361 (6400 July 2018). ISSN: 1095-9203. DOI: 10.1126/science.aar2555. ppublish.

[151] A. Boija, I. A. Klein, B. R. Sabari, A. Dall'Agnese, E. L. Coffey, A. V. Zamudio, C. H. Li, K. Shrinivas, J. C. Manteiga, N. M. Hannett, B. J. Abraham, L. K. Afeyan, Y. E. Guo, J. K. Rimel, C. B. Fant, J. Schuijers, T. I. Lee, D. J. Taatjes and R. A. Young. Transcription Factors Activate Genes through the Phase-

Separation Capacity of Their Activation Domains. *Cell* 175 (7 Dec. 2018), 1842–1855.e16. ISSN: 1097-4172. DOI: 10.1016/j.cell.2018.10.042. ppublish.

[152] K. Shrinivas, B. R. Sabari, E. L. Coffey, I. A. Klein, A. Boija, A. V. Zamudio, J. Schuijers, N. M. Hannett, P. A. Sharp, R. A. Young and A. K. Chakraborty. Enhancer Features that Drive Formation of Transcriptional Condensates. *Molecular cell* 75 (3 Aug. 2019), 549–561.e7. ISSN: 1097-4164. DOI: 10.1016/j.molcel.2019.07.009. ppublish.

[153] T. Di Domenico, I. Walsh, A. J. M. Martin and S. C. E. Tosatto. MobiDB: a comprehensive database of intrinsic protein disorder annotations. *Bioinformatics (Oxford, England)* 28 (15 Aug. 2012), 2080–2081. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/bts327. ppublish.

[154] E. Potenza, T. Di Domenico, I. Walsh and S. C. E. Tosatto. MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. *Nucleic acids research* 43 (Database issue Jan. 2015), D315–D320. ISSN: 1362-4962. DOI: 10.1093/nar/gku982. ppublish.

[155] D. Piovesan and S. C. E. Tosatto. Mobi 2.0: an improved method to define intrinsic disorder, mobility and linear binding regions in protein structures. *Bioinformatics (Oxford, England)* 34 (1 Jan. 2018), 122–123. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btx592. ppublish.

[156] C. Camilloni, A. De Simone, W. F. Vranken and M. Vendruscolo. Determination of secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts. *Biochemistry* 51 (11 Mar. 2012), 2224–2231. ISSN: 1520-4995. DOI: 10.1021/bi3001825. ppublish.

[157] M. V. Berjanskii and D. S. Wishart. A simple method to predict protein flexibility using secondary chemical shifts. *Journal of the American Chemical Society* 127 (43 Nov. 2005), 14970–14971. ISSN: 0002-7863. DOI: 10.1021/ja054842f. ppublish.

[158] M. Berjanskii and D. S. Wishart. NMR: prediction of protein flexibility. *Nature protocols* 1 (2 2006), 683–688. ISSN: 1750-2799. DOI: 10.1038/nprot.2006.108. ppublish.

[159] M. V. Berjanskii and D. S. Wishart. The RCI server: rapid and accurate calculation of protein flexibility using chemical shifts. *Nucleic acids research* 35 (Web Server issue July 2007), W531–W537. ISSN: 1362-4962. DOI: 10.1093/nar/gkm328. ppublish.

[160] Z. Dosztányi, V. Csizmok, P. Tompa and I. Simon. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics (Oxford, England)* 21 (16 Aug. 2005), 3433–3434. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bti541. ppublish.

[161] B. Mészáros, G. Erdos and Z. Dosztányi. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic acids research* 46 (W1 July 2018), W329–W337. ISSN: 1362-4962. DOI: 10.1093/nar/gky384. ppublish.

[162] Z. Dosztányi. Prediction of protein disorder based on IUPred. *Protein science : a publication of the Protein Society* 27 (1 Jan. 2018), 331–340. ISSN: 1469-896X. DOI: 10.1002/pro.3334. ppublish.

[163] B. Mészáros, I. Simon and Z. Dosztányi. Prediction of protein binding regions in disordered proteins. *PLoS computational biology* 5 (5 May 2009), e1000376. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1000376. ppublish.

[164] Z. Dosztányi, B. Mészáros and I. Simon. ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics* 25.20 (2009), 2745–2746.

[165] I. Walsh, A. J. M. Martin, T. Di Domenico and S. C. E. Tosatto. ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics (Oxford, England)* 28 (4 Feb. 2012), 503–509. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btr682. ppublish.

[166] D. Piovesan, I. Walsh, G. Minervini and S. C. E. Tosatto. FELLS: fast estimator of latent local structure. *Bioinformatics (Oxford, England)* 33 (12 June 2017), 1889–1891. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btx085. ppublish.

[167] D. Piovesan, G. Minervini and S. C. E. Tosatto. The RING 2.0 web server for high quality residue interaction networks. *Nucleic acids research* 44 (W1 July 2016), W367–W374. ISSN: 1362-4962. DOI: 10.1093/nar/gkw315. ppublish.

[168] R. Linding, L. J. Jensen, F. Diella, P. Bork, T. J. Gibson and R. B. Russell. Protein disorder prediction: implications for structural proteomics. *Structure (London, England : 1993)* 11 (11 Nov. 2003), 1453–1459. ISSN: 0969-2126. DOI: 10.1016/j.str.2003.10.002. ppublish.

[169] R. Linding, R. B. Russell, V. Neduva and T. J. Gibson. GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic acids research* 31 (13 July 2003), 3701–3708. ISSN: 1362-4962. DOI: 10.1093/nar/gkg519. ppublish.

[170] Z. R. Yang, R. Thomson, P. McNeil and R. M. Esnouf. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics (Oxford, England)* 21 (16 Aug. 2005), 3369–3376. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bti534. ppublish.

[171] K. Peng, P. Radivojac, S. Vucetic, A. K. Dunker and Z. Obradovic. Length-dependent prediction of protein intrinsic disorder. *BMC bioinformatics* 7 (Apr. 2006), 208. ISSN: 1471-2105. DOI: 10.1186/1471-2105-7-208. epublish.

[172] Z. Obradovic, K. Peng, S. Vucetic, P. Radivojac and A. K. Dunker. Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins* 61 Suppl 7 (2005), 176–182. ISSN: 1097-0134. DOI: 10.1002/prot.20735. ppublish.

[173] J. C. Wootton and S. Federhen. Analysis of compositionally biased regions in sequence databases. *Methods in enzymology* 266 (1996), 554–571. ISSN: 0076-6879. DOI: 10.1016/s0076-6879(96)66035-2. ppublish.

[174] D. T. Jones and M. B. Swindells. Getting the most from PSI-BLAST. *Trends in biochemical sciences* 27 (3 Mar. 2002), 161–164. ISSN: 0968-0004. DOI: 10.1016/s0968-0004(01)02039-4. ppublish.

[175] E. Cilia, R. Pancsa, P. Tompa, T. Lenaerts and W. F. Vranken. From protein sequence to dynamics and disorder with DynaMine. *Nature communications* 4 (2013), 2741. ISSN: 2041-1723. DOI: 10.1038/ncomms3741. ppublish.

[176] Z. Dosztanyi, V. Csizmok, P. Tompa and I. Simon. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *Journal of molecular biology* 347.4 (2005), 827–839.

[177] P. Baldi, S. Brunak, P. Frasconi, G. Soda and G. Pollastri. Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics* 15.11 (1999), 937–946.

[178] M. J. Mizianty, W. Stach, K. Chen, K. D. Kedarisetti, F. M. Disfani and L. Kurgan. Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. *Bioinformatics (Oxford, England)* 26 (18 Sept. 2010), i489–i496. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btq373. ppublish.

[179] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22 (12 Dec. 1983), 2577–2637. ISSN: 0006-3525. DOI: 10.1002/bip.360221211. ppublish.

[180] G. Deléage and B. Roux. An algorithm for protein secondary structure prediction based on class prediction. *Protein engineering* 1 (4 1987), 289–294. ISSN: 0269-2139. DOI: 10.1093/protein/1.4.289. ppublish.

[181] A. Andreeva, D. Howorth, C. Chothia, E. Kulesha and A. G. Murzin. SCOP2 prototype: a new approach to protein structure mining. *Nucleic acids research* 42.D1 (2014), D310–D314.

[182] A. Andreeva, E. Kulesha, J. Gough and A. G. Murzin. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic acids research* 48.D1 (2020), D376–D382.

[183] F. Sanger and A. R. Coulson. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology* 94 (3 May 1975), 441–448. ISSN: 0022-2836. DOI: 10.1016/0022-2836(75)90213-2. ppublish.

[184] E. S. Lander. Initial impact of the sequencing of the human genome. *Nature* 470 (7333 Feb. 2011), 187–197. ISSN: 1476-4687. DOI: 10.1038/nature09792. ppublish.

[185] I. H. G. S. Consortium. Finishing the euchromatic sequence of the human genome. *Nature* 431 (7011 Oct. 2004), 931–945. ISSN: 1476-4687. DOI: 10.1038/nature03001. ppublish.

[186] J. C. Venter et al. The sequence of the human genome. *Science (New York, N.Y.)* 291 (5507 Feb. 2001), 1304–1351. ISSN: 0036-8075. DOI: 10.1126/science.1058040. ppublish.

[187] P. Nyrén, B. Pettersson and M. Uhlén. Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. *Analytical biochemistry* 208 (1 Jan. 1993), 171–175. ISSN: 0003-2697. DOI: 10.1006/abio.1993.1024. ppublish.

[188] B. L. Karger and A. Guttman. DNA sequencing by CE. *Electrophoresis* 30 Suppl 1 (June 2009), S196–S202. ISSN: 1522-2683. DOI: 10.1002/elps.200900218. ppublish.

[189] B. Canard and R. S. Sarfati. DNA polymerase fluorescent substrates with reversible 3'-tags. *Gene* 148 (1 Oct. 1994), 1–6. ISSN: 0378-1119. DOI: 10.1016/0378-1119(94)90226-7. ppublish.

[190] A. Rhoads and K. F. Au. PacBio Sequencing and Its Applications. *Genomics, proteomics & bioinformatics* 13 (5 Oct. 2015), 278–289. ISSN: 2210-3244. DOI: 10.1016/j.gpb.2015.08.002. ppublish.

[191] V. Pandey, R. C. Nutter and E. Prediger. Applied biosystems solid™ system: ligation-based sequencing. *Next Generation Genome Sequencing: Towards Personalized Medicine* (2008), 29–42.

[192] J. Quick et al. Real-time, portable genome sequencing for Ebola surveillance. *Nature* 530 (7589 Feb. 2016), 228–232. ISSN: 1476-4687. DOI: 10.1038/nature16996. ppublish.

[193] A. A. Votintseva, P. Bradley, L. Pankhurst, C. Del Ojo Elias, M. Loose, K. Nilgiriwala, A. Chatterjee, E. G. Smith, N. Sanderson, T. M. Walker, M. R. Morgan, D. H. Wyllie, A. S. Walker, T. E. A. Peto, D. W. Crook and Z. Iqbal. Same-Day Diagnostic and Surveillance Data for Tuberculosis via Whole-Genome Sequencing of Direct Respiratory Samples. *Journal of clinical microbiology* 55 (5 May 2017), 1285–1298. ISSN: 1098-660X. DOI: 10.1128/JCM.02483-16. ppublish.

[194] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* 5 (7 July 2008), 621–628. ISSN: 1548-7105. DOI: 10.1038/nmeth.1226. ppublish.

[195] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein and M. Snyder. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science (New York, N.Y.)* 320 (5881 June 2008), 1344–1349. ISSN: 1095-9203. DOI: 10.1126/science.1158441. ppublish.

[196] R. Lister, R. C. O'Malley, J. Tonti-Filippini, B. D. Gregory, C. C. Berry, A. H. Millar and J. R. Ecker. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* 133 (3 May 2008), 523–536. ISSN: 1097-4172. DOI: 10.1016/j.cell.2008.03.029. ppublish.

[197] Y. Liao and W. Shi. Read trimming is not required for mapping and quantification of RNA-seq reads. *bioRxiv* (2019). DOI: 10.1101/833962. eprint: https://www.biorxiv.org/content/early/2019/11/07/833962.full.pdf. URL: https://www.biorxiv.org/content/early/2019/11/07/833962.

[198] M. R. Friedländer, S. D. Mackowiak, N. Li, W. Chen and N. Rajewsky. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic acids research* 40 (1 Jan. 2012), 37–52. ISSN: 1362-4962. DOI: 10.1093/nar/gkr688. ppublish.

[199] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome biology* 11 (10 2010), R106. ISSN: 1474-760X. DOI: 10.1186/gb-2010-11-10-r106. ppublish.

[200] M. I. Love, W. Huber and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* 15 (12 2014), 550. ISSN: 1474-760X. DOI: 10.1186/s13059-014-0550-8. ppublish.

[201] M. D. Robinson, D. J. McCarthy and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)* 26 (1 Jan. 2010), 139–140. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp616. ppublish.

[202] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi and G. K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research* 43 (7 Apr. 2015), e47. ISSN: 1362-4962. DOI: 10.1093/nar/gkv007. ppublish.

[203] N. L. Bray, H. Pimentel, P. Melsted and L. Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology* 34 (5 May 2016), 525–527. ISSN: 1546-1696. DOI: 10.1038/nbt.3519. ppublish.

[204] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry and C. Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods* 14 (4 Apr. 2017), 417–419. ISSN: 1548-7105. DOI: 10.1038/nmeth.4197. ppublish.

[205] P. J. Park. ChIP-seq: advantages and challenges of a maturing technology. *Nature reviews. Genetics* 10 (10 Oct. 2009), 669–680. ISSN: 1471-0064. DOI: 10.1038/nrg2641. ppublish.

[206] R. Nakato and T. Sakata. Methods for ChIP-seq analysis: A practical workflow and advanced applications. *Methods (San Diego, Calif.)* (Mar. 2020). ISSN: 1095-9130. DOI: 10.1016/j.ymeth.2020.03.005. aheadofprint.

[207] C. A. Davis, B. C. Hitz, C. A. Sloan, E. T. Chan, J. M. Davidson, I. Gabdank, J. A. Hilton, K. Jain, U. K. Baymuradov, A. K. Narayanan, K. C. Onate, K. Graham, S. R. Miyasato, T. R. Dreszer, J. S. Strattan, O. Jolanki, F. Y. Tanaka and J. M. Cherry. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic acids research* 46 (D1 Jan. 2018), D794–D801. ISSN: 1362-4962. DOI: 10.1093/nar/gkx1081. ppublish.

[208] R. E. Consortium, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y.-C. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K.-H. Farh, S. Feizi, R. Karlic, A.-R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. M. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L.-H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang and M. Kellis. Integrative analysis

of 111 reference human epigenomes. *Nature* 518 (7539 Feb. 2015), 317–330. ISSN: 1476-4687. DOI: 10.1038/nature14248. ppublish.

[209] P. J. Skene and S. Henikoff. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife* 6 (Jan. 2017). ISSN: 2050-084X. DOI: 10.7554/eLife.21856. epublish.

[210] H. S. Rhee and B. F. Pugh. ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. *Current protocols in molecular biology* Chapter 21 (Oct. 2012), Unit 21.24. ISSN: 1934-3647. DOI: 10.1002/0471142727.mb2124s100. ppublish.

[211] T. R. Riley, M. Slattery, N. Abe, C. Rastogi, D. Liu, R. S. Mann and H. J. Bussemaker. SELEX-seq: a method for characterizing the complete repertoire of binding site preferences for transcription factor complexes. *Methods in molecular biology (Clifton, N.J.)* 1196 (2014), 255–278. ISSN: 1940-6029. DOI: 10.1007/978-1-4939-1242-1_16. ppublish.

[212] J. Dekker, K. Rippe, M. Dekker and N. Kleckner. Capturing chromosome conformation. *Science (New York, N.Y.)* 295 (5558 Feb. 2002), 1306–1311. ISSN: 1095-9203. DOI: 10.1126/science.1067799. ppublish.

[213] Z. Zhao, G. Tavoosidana, M. Sjölinder, A. Göndör, P. Mariano, S. Wang, C. Kanduri, M. Lezcano, K. S. Sandhu, U. Singh, V. Pant, V. Tiwari, S. Kurukuti and R. Ohlsson. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature genetics* 38 (11 Nov. 2006), 1341–1347. ISSN: 1061-4036. DOI: 10.1038/ng1891. ppublish.

[214] M. Simonis, P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. de Wit, B. van Steensel and W. de Laat. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature genetics* 38 (11 Nov. 2006), 1348–1354. ISSN: 1061-4036. DOI: 10.1038/ng1896. ppublish.

[215] J. Dostie, T. A. Richmond, R. A. Arnaout, R. R. Selzer, W. L. Lee, T. A. Honan, E. D. Rubio, A. Krumm, J. Lamb, C. Nusbaum, R. D. Green and J. Dekker. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome*

research 16 (10 Oct. 2006), 1299–1309. ISSN: 1088-9051. DOI: 10.1101/gr.5571506. ppublish.

[216] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander and J. Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York, N.Y.)* 326 (5950 Oct. 2009), 289–293. ISSN: 1095-9203. DOI: 10.1126/science.1181369. ppublish.

[217] J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang and W. J. Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods* 10 (12 Dec. 2013), 1213–1218. ISSN: 1548-7105. DOI: 10.1038/nmeth.2688. ppublish.

[218] M. R. Corces, J. M. Granja, S. Shams, B. H. Louie, J. A. Seoane, W. Zhou, T. C. Silva, C. Groeneveld, C. K. Wong, S. W. Cho, A. T. Satpathy, M. R. Mumbach, K. A. Hoadley, A. G. Robertson, N. C. Sheffield, I. Felau, M. A. A. Castro, B. P. Berman, L. M. Staudt, J. C. Zenklusen, P. W. Laird, C. Curtis, C. G. A. A. Network, W. J. Greenleaf and H. Y. Chang. The chromatin accessibility landscape of primary human cancers. *Science (New York, N.Y.)* 362 (6413 Oct. 2018). ISSN: 1095-9203. DOI: 10.1126/science.aav1898. ppublish.

[219] A. Dash, I. P. Maine, S. Varambally, R. Shen, A. M. Chinnaiyan and M. A. Rubin. Changes in differential gene expression because of warm ischemia time of radical prostatectomy specimens. *The American journal of pathology* 161 (5 Nov. 2002), 1743–1748. ISSN: 0002-9440. DOI: 10.1016/S0002-9440(10)64451-3. ppublish.

[220] M. Annala, K. Kivinummi, J. Tuominen, S. Karakurt, K. Granberg, L. Latonen, A. Ylipää, L. Sjöblom, P. Ruusuvuori, O. Saramäki, K. M. Kaukoniemi, O. Yli-Harja, R. L. Vessella, T. L. J. Tammela, W. Zhang, T. Visakorpi and M. Nykter. Recurrent SKIL-activating rearrangements in ETS-negative prostate cancer. *Oncotarget* 6 (8 Mar. 2015), 6235–6250. ISSN: 1949-2553. DOI: 10.18632/oncotarget.3359. ppublish.

[221] A. Ylipää, K. Kivinummi, A. Kohvakka, M. Annala, L. Latonen, M. Scaravilli, K. Kartasalo, S.-P. Leppänen, S. Karakurt, J. Seppälä, O. Yli-Harja, T. L. J. Tammela, W. Zhang, T. Visakorpi and M. Nykter. Transcriptome Sequencing Reveals PCAT5 as a Novel ERG-Regulated Long Noncoding RNA in Prostate Cancer. *Cancer research* 75 (19 Oct. 2015), 4026–4031. ISSN: 1538-7445. DOI: 10.1158/0008-5472.CAN-15-0217. ppublish.

[222] L. Latonen, E. Afyounian, A. Jylhä, J. Nättinen, U. Aapola, M. Annala, K. K. Kivinummi, T. T. L. Tammela, R. W. Beuerman, H. Uusitalo, M. Nykter and T. Visakorpi. Integrative proteomics in prostate cancer uncovers robustness against genomic and transcriptomic aberrations during disease progression. *Nature communications* 9 (1 Mar. 2018), 1176. ISSN: 2041-1723. DOI: 10.1038/s41467-018-03573-6. epublish.

[223] O. R. Saramäki, A. E. Harjula, P. M. Martikainen, R. L. Vessella, T. L. J. Tammela and T. Visakorpi. TMPRSS2:ERG fusion identifies a subgroup of prostate cancers with a favorable prognosis. *Clinical cancer research : an official journal of the American Association for Cancer Research* 14 (11 June 2008), 3395–3400. ISSN: 1078-0432. DOI: 10.1158/1078-0432.CCR-07-2051. ppublish.

[224] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson and T. R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)* 29 (1 Jan. 2013), 15–21. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/bts635. ppublish.

[225] A. Kozomara, M. Birgaoanu and S. Griffiths-Jones. miRBase: from microRNA sequences to function. *Nucleic acids research* 47.D1 (2019), D155–D162. DOI: 10.1093/nar/gky1141.

[226] S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh and C. K. Glass. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell* 38 (4 May 2010), 576–589. ISSN: 1097-4164. DOI: 10.1016/j.molcel.2010.05.004. ppublish.

[227] E. P. Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489 (7414 Sept. 2012), 57–74. ISSN: 1476-4687. DOI: 10.1038/nature11247. ppublish.

[228] A. Pombo and N. Dillon. Three-dimensional genome architecture: players and mechanisms. *Nature reviews. Molecular cell biology* 16 (4 Apr. 2015), 245–257. ISSN: 1471-0080. DOI: 10.1038/nrm3965. ppublish.

[229] I. Yevshin, R. Sharipov, T. Valeev, A. Kel and F. Kolpakov. GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. *Nucleic acids research* 45 (D1 Jan. 2017), D61–D67. ISSN: 1362-4962. DOI: 10.1093/nar/gkw951. ppublish.

[230] U. Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic acids research* 47 (D1 Jan. 2019), D506–D515. ISSN: 1362-4962. DOI: 10.1093/nar/gky1049. ppublish.

[231] *Superheroic JavaScript MVW Framework*. URL: https://angularjs.org/.

[232] L.-H. Jiang, H.-d. Zhang and J.-H. Tang. MiR-30a: A Novel Biomarker and Potential Therapeutic Target for Cancer. *Journal of oncology* 2018 (2018), 5167829. ISSN: 1687-8450. DOI: 10.1155/2018/5167829. epublish.

[233] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* 25 (1 May 2000), 25–29. ISSN: 1061-4036. DOI: 10.1038/75556. ppublish.

[234] T. G. O. Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic acids research* 47 (D1 Jan. 2019), D330–D338. ISSN: 1362-4962. DOI: 10.1093/nar/gky1055. ppublish.

[235] T. UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic acids research* 46 (5 Mar. 2018), 2699. ISSN: 1362-4962. DOI: 10.1093/nar/gky092. ppublish.

[236] S. El-Gebali, J. Mistry, A. Bateman, S. R. Eddy, A. Luciani, S. C. Potter, M. Qureshi, L. J. Richardson, G. A. Salazar, A. Smart, E. L. L. Sonnhammer, L. Hirsh, L. Paladin, D. Piovesan, S. C. E. Tosatto and R. D. Finn. The Pfam protein families database in 2019. *Nucleic acids research* 47 (D1 Jan. 2019), D427–D432. ISSN: 1362-4962. DOI: 10.1093/nar/gky995. ppublish.

[237] D. Goodsell. Nucleosome. *RCSB Protein Data Bank* (July 2000). DOI: 10.2210/rcsb_pdb/mom_2000_7.

[238] D. S. Goodsell, S. Dutta, C. Zardecki, M. Voigt, H. M. Berman and S. K. Burley. The RCSB PDB "Molecule of the Month": Inspiring a Molecular View of Biology. *PLoS biology* 13 (5 May 2015), e1002140. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1002140. epublish.

[239] R. D. Page. GeneTree: comparing gene and species phylogenies using reconciled trees. *Bioinformatics (Oxford, England)* 14 (9 1998), 819–820. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/14.9.819. ppublish.

[240] M. Miskei, C. Antal and M. Fuxreiter. FuzDB: database of fuzzy complexes, a tool to develop stochastic structure-function relationships for protein complexes and higher-order assemblies. *Nucleic acids research* 45 (D1 Jan. 2017), D228–D235. ISSN: 1362-4962. DOI: 10.1093/nar/gkw1019. ppublish.

[241] M. Kumar, M. Gouw, S. Michael, H. Sámano-Sánchez, R. Pancsa, J. Glavina, A. Diakogianni, J. A. Valverde, D. Bukirova, J. Čalyševa, N. Palopoli, N. E. Davey, L. B. Chemes and T. J. Gibson. ELM-the eukaryotic linear motif resource in 2020. *Nucleic acids research* 48 (D1 Jan. 2020), D296–D306. ISSN: 1362-4962. DOI: 10.1093/nar/gkz1030. ppublish.

[242] S. Fukuchi, S. Sakamoto, Y. Nobe, S. D. Murakami, T. Amemiya, K. Hosoda, R. Koike, H. Hiroaki and M. Ota. IDEAL: Intrinsically Disordered proteins with Extensive Annotations and Literature. *Nucleic acids research* 40 (Database issue Jan. 2012), D507–D511. ISSN: 1362-4962. DOI: 10.1093/nar/gkr884. ppublish.

[243] S. Fukuchi, T. Amemiya, S. Sakamoto, Y. Nobe, K. Hosoda, Y. Kado, S. D. Murakami, R. Koike, H. Hiroaki and M. Ota. IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners. *Nucleic acids research* 42 (Database issue Jan. 2014), D320–D325. ISSN: 1362-4962. DOI: 10.1093/nar/gkt1010. ppublish.

[244] E. Schad, E. Fichó, R. Pancsa, I. Simon, Z. Dosztányi and B. Mészáros. DIBS: a repository of disordered binding sites mediating interactions with ordered proteins. *Bioinformatics (Oxford, England)* 34 (3 Feb. 2018), 535–537. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btx640. ppublish.

[245] E. Fichó, I. Reményi, I. Simon and B. Mészáros. MFIB: a repository of protein complexes with mutual folding induced by binding. *Bioinformatics (Oxford, England)* 33 (22 Nov. 2017), 3682–3684. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btx486. ppublish.

[246] T. E. Lewis, I. Sillitoe, N. Dawson, S. D. Lam, T. Clarke, D. Lee, C. Orengo and J. Lees. Gene3D: Extensive prediction of globular domains in proteins. *Nucleic acids research* 46 (D1 Jan. 2018), D435–D439. ISSN: 1362-4962. DOI: 10.1093/nar/gkx1069. ppublish.

[247] A. M. Monzon, C. O. Rohr, M. S. Fornasari and G. Parisi. CoDNaS 2.0: a comprehensive database of protein conformational diversity in the native state. *Database : the journal of biological databases and curation* 2016 (2016). ISSN: 1758-0463. DOI: 10.1093/database/baw038. epublish.

[248] A. M. Monzon, E. Juritz, M. S. Fornasari and G. Parisi. CoDNaS: a database of conformational diversity in the native state of proteins. *Bioinformatics (Oxford, England)* 29 (19 Oct. 2013), 2512–2514. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btt405. ppublish.

[249] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne. The Protein Data Bank. *Nucleic acids research* 28 (1 Jan. 2000), 235–242. ISSN: 0305-1048. DOI: 10.1093/nar/28.1.235. ppublish.

[250] S. K. Burley, H. M. Berman, C. Bhikadiya, C. Bi, L. Chen, L. Di Costanzo, C. Christie, K. Dalenberg, J. M. Duarte, S. Dutta, Z. Feng, S. Ghosh, D. S. Goodsell, R. K. Green, V. Guranovic, D. Guzenko, B. P. Hudson, T. Kalro, Y. Liang, R. Lowe, H. Namkoong, E. Peisach, I. Periskova, A. Prlic, C. Randle, A. Rose, P. Rose, R. Sala, M. Sekharan, C. Shao, L. Tan, Y.-P. Tao, Y. Valasatava, M. Voigt, J. Westbrook, J. Woo, H. Yang, J. Young, M. Zhuravleva and C. Zardecki. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic acids research* 47 (D1 Jan. 2019), D464–D474. ISSN: 1362-4962. DOI: 10.1093/nar/gky1004. ppublish.

[251] E. L. Ulrich, H. Akutsu, J. F. Doreleijers, Y. Harano, Y. E. Ioannidis, J. Lin, M. Livny, S. Mading, D. Maziuk, Z. Miller, E. Nakatani, C. F. Schulte, D. E. Tolmie, R. Kent Wenger, H. Yao and J. L. Markley. BioMagResBank. *Nucleic*

acids research 36 (Database issue Jan. 2008), D402–D408. ISSN: 1362-4962. DOI: 10.1093/nar/gkm957. ppublish.

[252] M. Necci, D. Piovesan, Z. Dosztányi and S. C. E. Tosatto. MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins. *Bioinformatics (Oxford, England)* 33 (9 May 2017), 1402–1404. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btx015. ppublish.

[253] B. Li, M. Carey and J. L. Workman. The role of chromatin during transcription. *Cell* 128 (4 Feb. 2007), 707–719. ISSN: 0092-8674. DOI: 10.1016/j.cell.2007.01.015. ppublish.

[254] Y. Liu, A. Beyer and R. Aebersold. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* 165 (3 Apr. 2016), 535–550. ISSN: 1097-4172. DOI: 10.1016/j.cell.2016.03.014. ppublish.

[255] P. Rajbhandari, B. J. Thomas, A.-C. Feng, C. Hong, J. Wang, L. Vergnes, T. Sallam, B. Wang, J. Sandhu, M. M. Seldin, A. J. Lusis, L. G. Fong, M. Katz, R. Lee, S. G. Young, K. Reue, S. T. Smale and P. Tontonoz. IL-10 Signaling Remodels Adipose Chromatin Architecture to Limit Thermogenesis and Energy Expenditure. *Cell* 172 (1-2 Jan. 2018), 218–233.e17. ISSN: 1097-4172. DOI: 10.1016/j.cell.2017.11.019. ppublish.

[256] C. D. Scharer, B. G. Barwick, M. Guo, A. P. R. Bally and J. M. Boss. Plasma cell differentiation is controlled by multiple cell division-coupled epigenetic programs. *Nature communications* 9 (1 Apr. 2018), 1698. ISSN: 2041-1723. DOI: 10.1038/s41467-018-04125-8. epublish.

[257] C. G. Toenhake, S. A.-K. Fraschka, M. S. Vijayabaskar, D. R. Westhead, S. J. van Heeringen and R. Bártfai. Chromatin Accessibility-Based Characterization of the Gene Regulatory Network Underlying Plasmodium falciparum Blood-Stage Development. *Cell host & microbe* 23 (4 Apr. 2018), 557–569.e9. ISSN: 1934-6069. DOI: 10.1016/j.chom.2018.03.007. ppublish.

[258] J. Wu, J. Xu, B. Liu, G. Yao, P. Wang, Z. Lin, B. Huang, X. Wang, T. Li, S. Shi, N. Zhang, F. Duan, J. Ming, X. Zhang, W. Niu, W. Song, H. Jin, Y. Guo, S. Dai, L. Hu, L. Fang, Q. Wang, Y. Li, W. Li, J. Na, W. Xie and Y. Sun. Chromatin analysis in human early development reveals epigenetic transition during ZGA. *Nature* 557 (7704 May 2018), 256–260. ISSN: 1476-4687. DOI: 10.1038/s41586-018-0080-8. ppublish.

[259] E. Markenscoff-Papadimitriou, W. E. Allen, B. M. Colquitt, T. Goh, K. K. Murphy, K. Monahan, C. P. Mosley, N. Ahituv and S. Lomvardas. Enhancer interaction networks as a means for singular olfactory receptor expression. *Cell* 159 (3 Oct. 2014), 543–557. ISSN: 1097-4172. DOI: 10.1016/j.cell.2014.09.033. ppublish.

[260] E. Ing-Simmons, V. C. Seitan, A. J. Faure, P. Flicek, T. Carroll, J. Dekker, A. G. Fisher, B. Lenhard and M. Merkenschlager. Spatial enhancer clustering and regulation of enhancer-proximal genes by cohesin. *Genome research* 25 (4 Apr. 2015), 504–513. ISSN: 1549-5469. DOI: 10.1101/gr.184986.114. ppublish.

[261] R. A. Beagrie, A. Scialdone, M. Schueler, D. C. A. Kraemer, M. Chotalia, S. Q. Xie, M. Barbieri, I. de Santiago, L.-M. Lavitas, M. R. Branco, J. Fraser, J. Dostie, L. Game, N. Dillon, P. A. W. Edwards, M. Nicodemi and A. Pombo. Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* 543 (7646 Mar. 2017), 519–524. ISSN: 1476-4687. DOI: 10.1038/nature21411. ppublish.

[262] D. Y. Takeda, S. Spisák, J.-H. Seo, C. Bell, E. O'Connor, K. Korthauer, D. Ribli, I. Csabai, N. Solymosi, Z. Szállási, D. R. Stillman, P. Cejas, X. Qiu, H. W. Long, V. Tisza, P. V. Nuzzo, M. Rohanizadegan, M. M. Pomerantz, W. C. Hahn and M. L. Freedman. A Somatically Acquired Enhancer of the Androgen Receptor Is a Noncoding Driver in Advanced Prostate Cancer. *Cell* 174 (2 July 2018), 422–432.e13. ISSN: 1097-4172. DOI: 10.1016/j.cell.2018.05.037. ppublish.

[263] M. M. Pomerantz, F. Li, D. Y. Takeda, R. Lenci, A. Chonkar, M. Chabot, P. Cejas, F. Vazquez, J. Cook, R. A. Shivdasani, M. Bowden, R. Lis, W. C. Hahn, P. W. Kantoff, M. Brown, M. Loda, H. W. Long and M. L. Freedman. The androgen receptor cistrome is extensively reprogrammed in human prostate tumorigenesis. *Nature genetics* 47 (11 Nov. 2015), 1346–1351. ISSN: 1546-1718. DOI: 10.1038/ng.3419. ppublish.

[264] P. R. Braadland and A. Urbanucci. Chromatin reprogramming as an adaptation mechanism in advanced prostate cancer. *Endocrine-related cancer* 26 (4 Apr. 2019), R211–R235. ISSN: 1479-6821. DOI: 10.1530/ERC-18-0579. ppublish.

[265] J. D. Norris, C.-Y. Chang, B. M. Wittmann, R. S. Kunder, H. Cui, D. Fan, J. D. Joseph and D. P. McDonnell. The homeodomain protein HOXB13 regulates

the cellular response to androgens. *Molecular cell* 36 (3 Nov. 2009), 405–416. ISSN: 1097-4164. DOI: 10.1016/j.molcel.2009.10.020. ppublish.

[266] C. M. Ewing, A. M. Ray, E. M. Lange, K. A. Zuhlke, C. M. Robbins, W. D. Tembe, K. E. Wiley, S. D. Isaacs, D. Johng, Y. Wang, C. Bizon, G. Yan, M. Gielzak, A. W. Partin, V. Shanmugam, T. Izatt, S. Sinari, D. W. Craig, S. L. Zheng, P. C. Walsh, J. E. Montie, J. Xu, J. D. Carpten, W. B. Isaacs and K. A. Cooney. Germline mutations in HOXB13 and prostate-cancer risk. *The New England journal of medicine* 366 (2 Jan. 2012), 141–149. ISSN: 1533-4406. DOI: 10.1056/NEJMoa1110000. ppublish.

[267] H. Yuan, A. Gong and C. Y. F. Young. Involvement of transcription factor Sp1 in quercetin-mediated inhibitory effect on the androgen receptor in human prostate cancer cells. *Carcinogenesis* 26 (4 Apr. 2005), 793–801. ISSN: 0143-3334. DOI: 10.1093/carcin/bgi021. ppublish.

[268] P.-H. Chen, Y.-P. Tsao, C.-C. Wang and S.-L. Chen. Nuclear receptor interaction protein, a coactivator of androgen receptors (AR), is regulated by AR and Sp1 to feed forward and activate its own gene expression through AR protein stability. *Nucleic acids research* 36 (1 Jan. 2008), 51–66. ISSN: 1362-4962. DOI: 10.1093/nar/gkm942. ppublish.

[269] U. T. Sankpal, S. Goodison, M. Abdelrahim and R. Basha. Targeting Sp1 transcription factors in prostate cancer therapy. *Medicinal chemistry (Shariqah (United Arab Emirates))* 7 (5 Sept. 2011), 518–525. ISSN: 1875-6638. DOI: 10.2174/157340611796799203. ppublish.

[270] N. H. Farina, A. Zingiryan, J. A. Akech, C. J. Callahan, H. Lu, J. L. Stein, L. R. Languino, G. S. Stein and J. B. Lian. A microRNA/Runx1/Runx2 network regulates prostate tumor progression from onset to adenocarcinoma in TRAMP mice. *Oncotarget* 7.43 (Oct. 2016), 70462–70474.

[271] K. Takayama, T. Suzuki, S. Tsutsumi, T. Fujimura, T. Urano, S. Takahashi, Y. Homma, H. Aburatani and S. Inoue. RUNX1, an androgen- and EZH2-regulated gene, has differential roles in AR-dependent and -independent prostate cancer. *Oncotarget* 6.4 (Feb. 2015), 2263–2276.

[272] C. M. Koh, C. J. Bieberich, C. V. Dang, W. G. Nelson, S. Yegnasubramanian and A. M. De Marzo. MYC and Prostate Cancer. *Genes Cancer* 1.6 (June 2010), 617–628.

[273] C. Elix, S. K. Pal and J. O. Jones. The role of peroxisome proliferator-activated receptor gamma in prostate cancer. *Asian J Androl* 20.3 (2018), 238–243.

[274] M. Puhr, J. Hoefer, A. Eigentler, C. Ploner, F. Handle, G. Schaefer, J. Kroon, A. Leo, I. Heidegger, I. Eder, Z. Culig, G. Van der Pluijm and H. Klocker. The Glucocorticoid Receptor Is a Key Player for Prostate Cancer Cell Survival and a Target for Improved Antiandrogen Therapy. *Clin Cancer Res* 24.4 (Feb. 2018), 927–938.

[275] J. W. Park, J. K. Lee, O. N. Witte and J. Huang. FOXA2 is a sensitive and specific marker for small cell neuroendocrine carcinoma of the prostate. *Mod Pathol* 30.9 (Sept. 2017), 1262–1272.

[276] I. Yevshin, R. Sharipov, S. Kolmykov, Y. Kondrakhin and F. Kolpakov. GTRD: a database on gene transcription regulation-2019 update. *Nucleic acids research* 47 (D1 Jan. 2019), D100–D105. ISSN: 1362-4962. DOI: 10.1093/nar/gky1128. ppublish.

[277] E. Hibino, R. Inoue, M. Sugiyama, J. Kuwahara, K. Matsuzaki and M. Hoshino. Interaction between intrinsically disordered regions in transcription factors Sp1 and TAF4. *Protein science : a publication of the Protein Society* 25 (11 Nov. 2016), 2006–2017. ISSN: 1469-896X. DOI: 10.1002/pro.3013. ppublish.

[278] J. Uusi-Mäkelä*, E. Afyounian*, F. Tabaro*, T. Häkkinen*, A. Lussana, A. Shcherban, M. Annala, R. Nurminen, K. Kivinummi, T. L. Tammela, A. Urbanucci, L. Latonen, J. Kesseli, K. J. Granberg, T. Visakorpi and M. Nykter. Chromatin accessibility analysis uncovers regulatory element landscape in prostate cancer progression. *bioRxiv* (2020). DOI: 10.1101/2020.09.08.287268. eprint: https://www.biorxiv.org/content/early/2020/09/09/2020.09.08.287268.full.pdf. URL: https://www.biorxiv.org/content/early/2020/09/09/2020.09.08.287268.

[279] D. Piovesan*, F. Tabaro*, I. Mičetić, M. Necci, F. Quaglia, C. J. Oldfield, M. C. Aspromonte, N. E. Davey, R. Davidović, Z. Dosztányi, A. Elofsson, A. Gasparini, A. Hatos, A. V. Kajava, L. Kalmar, E. Leonardi, T. Lazar, S. Macedo-Ribeiro, M. Macossay-Castillo, A. Meszaros, G. Minervini, N. Murvai, J. Pujols, D. B. Roche, E. Salladini, E. Schad, A. Schramm, B. Szabo, A. Tantos, F. Tonello, K. D. Tsirigos, N. Veljković, S. Ventura, W. Vranken, P. Warholm, V. N. Uversky, A. K. Dunker, S. Longhi, P. Tompa and S. C. E. Tosatto. DisProt

7.0: a major update of the database of disordered proteins. *Nucleic acids research* 45 (D1 Jan. 2017), D219–D227. ISSN: 1362-4962. DOI: 10.1093/nar/gkw1056. ppublish.

[280] D. Piovesan*, F. Tabaro*, L. Paladin, M. Necci, I. Micetic, C. Camilloni, N. Davey, Z. Dosztányi, B. Mészáros, A. M. Monzon, G. Parisi, E. Schad, P. Sormanni, P. Tompa, M. Vendruscolo, W. F. Vranken and S. C. E. Tosatto. MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins. *Nucleic acids research* 46 (D1 Jan. 2018), D471–D476. ISSN: 1362-4962. DOI: 10.1093/nar/gkx1071. ppublish.

PUBLICATIONS

PUBLICATION

I

Chromatin accessibility analysis uncovers regulatory element landscape in prostate cancer progression

J. Uusi-Mäkelä*, E. Afyounian*, F. Tabaro*, T. Häkkinen*, A. Lussana, A. Shcherban, M. Annala, R. Nurminen, K. Kivinummi, T. L. Tammela, A. Urbanucci, L. Latonen, J. Kesseli, K. J. Granberg, T. Visakorpi and M. Nykter

bioRxiv (2020)
DOI: 10.1101/2020.09.08.287268

PUBLICATION
II

DisProt 7.0: a major update of the database of disordered proteins.

D. Piovesan*, F. Tabaro*, I. Mičetić, M. Necci, F. Quaglia, C. J. Oldfield, M. C. Aspromonte, N. E. Davey, R. Davidović, Z. Dosztányi, A. Elofsson, A. Gasparini, A. Hatos, A. V. Kajava, L. Kalmar, E. Leonardi, T. Lazar, S. Macedo-Ribeiro, M. Macossay-Castillo, A. Meszaros, G. Minervini, N. Murvai, J. Pujols, D. B. Roche, E. Salladini, E. Schad, A. Schramm, B. Szabo, A. Tantos, F. Tonello, K. D. Tsirigos, N. Veljković, S. Ventura, W. Vranken, P. Warholm, V. N. Uversky, A. K. Dunker, S. Longhi, P. Tompa and S. C. E. Tosatto

Nucleic acids research 45.(2017), D219–D227
DOI: 10.1093/nar/gkw1056

Published online 24 November 2016 *Nucleic Acids Research, 2017, Vol. 45, Database issue* ***D219–D227***
doi: 10.1093/nar/gkw1056

DisProt 7.0: a major update of the database of disordered proteins

Damiano Piovesan[1,†], Francesco Tabaro[1,2,†], Ivan Mičetić[1], Marco Necci[1], Federica Quaglia[1], Christopher J. Oldfield[3], Maria Cristina Aspromonte[4], Norman E. Davey[5,6], Radoslav Davidović[7], Zsuzsanna Dosztányi[8,9], Arne Elofsson[10], Alessandra Gasparini[4], András Hatos[1,9], Andrey V. Kajava[11,12,13], Lajos Kalmar[9,14], Emanuela Leonardi[4], Tamas Lazar[15,16], Sandra Macedo-Ribeiro[17], Mauricio Macossay-Castillo[15,16], Attila Meszaros[9], Giovanni Minervini[1], Nikoletta Murvai[9], Jordi Pujols[18], Daniel B. Roche[11,12], Edoardo Salladini[19], Eva Schad[9], Antoine Schramm[19], Beata Szabo[9], Agnes Tantos[9], Fiorella Tonello[1,20], Konstantinos D. Tsirigos[10], Nevena Veljković[7], Salvador Ventura[18], Wim Vranken[15,16,21], Per Warholm[10], Vladimir N. Uversky[22,23], A. Keith Dunker[3], Sonia Longhi[19,*], Peter Tompa[9,15,16,*] and Silvio C.E. Tosatto[1,20,*]

[1]Department of Biomedical Sciences, University of Padova, I-35121 Padova, Italy, [2]Institute of Biosciences and Medical Technology, University of Tampere, Finland, [3]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, 46202 Indianapolis, IN, USA, [4]Department of Woman and Child Health, University of Padova, I-35128 Padova, Italy, [5]Conway Institute of Biomolecular & Biomedical Research, University College Dublin, Belfield, Dublin 4, Ireland, [6]Ireland UCD School of Medicine & Medical Science, University College Dublin, Belfield, Dublin 4, Ireland, [7]Centre for Multidisciplinary Research, Institute of Nuclear Sciences Vinca, University of Belgrade, 11001 Belgrade, Serbia, [8]MTA-ELTE Lendület Bioinformatics Research Group, Department of Biochemistry, Eötvös Loránd University, 1/c Pázmány Péter sétány, 1117 Budapest, Hungary, [9]Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, PO Box 7,H-1518 Budapest, Hungary, [10]Department of Biochemistry and Biophysics and Science for Life Laboratory, Stockholm University, Box 1031, 17121 Solna, Sweden, [11]Centre de Recherche en Biologie cellulaire de Montpellier (CRBM), UMR 5237 CNRS, Université Montpellier 1919 Route de Mende, Cedex 5, Montpellier 34293, France, [12]Institut de Biologie Computationnelle (IBC), Montpellier 34095, France, [13]University ITMO, Institute of Bioengineering, St. Petersburg 197101, Russia, [14]Department of Veterinary Medicine, University of Cambridge, Madingley Road, Cambridge CB3 0ES, UK, [15]Structural Biology Brussels, Vrije Universiteit Brussel (VUB), Brussels 1050, Belgium, [16]Structural Biology Research Center (SBRC), Flanders Institute for Biotechnology (VIB), Brussels 1050, Belgium, [17]Biomolecular Structure and Function Group, Instituto de Biologia Molecular e Celular (IBMC) and Instituto de Investigação e Inovação em Saúde (i3S), Universidade do Porto, 4200-135 Porto, Portugal, [18]Departament de Bioquimica i Biologia Molecular and Institut de Biotecnologia i Biomedicina, Universitat Autònoma de Barcelona, Bellaterra 08193, Spain, [19]Aix-Marseille Univ, CNRS, AFMB, UMR 7257, Marseille, France, [20]CNR Institute of Neurosceince, I-35121 Padova, Italy, [21]Interuniversity Institute of Bioinformatics in Brussels (IB2), ULB-VUB, Brussels 1050, Belgium, [22]Laboratory of Structural Dynamics, Stability and Folding of Proteins, Institute of Cytology, Russian Academy of Sciences, 194064 St. Petersburg, Russia and [23]Department of Molecular Medicine and USF Health Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida, Tampa, FL 33612, USA

Received September 27, 2016; Revised October 19, 2016; Editorial Decision October 20, 2016; Accepted October 21, 2016

[*]To whom correspondence should be addressed. Tel: +39 049 827 6269; Email: silvio.tosatto@unipd.it
Correspondence may also be addressed to Sonia Longhi. Tel: +33 4 91 82 55 80; Email: Sonia.Longhi@afmb.univ-mrs.fr
Correspondence may also be addressed to Peter Tompa. Tel: +32 2 629 1962; Email: ptompa@vub.ac.be
[†]These authors contributed equally to the paper as first authors.

ABSTRACT

The Database of Protein Disorder (DisProt, URL: www.disprot.org) has been significantly updated and upgraded since its last major renewal in 2007. The current release holds information on more than 800 entries of IDPs/IDRs, i.e. intrinsically disordered proteins or regions that exist and function without a well-defined three-dimensional structure. We have re-curated previous entries to purge DisProt from conflicting cases, and also upgraded the functional classification scheme to reflect continuous advance in the field in the past 10 years or so. We define IDPs as proteins that are disordered along their entire sequence, i.e. entirely lack structural elements, and IDRs as regions that are at least five consecutive residues without well-defined structure. We base our assessment of disorder strictly on experimental evidence, such as X-ray crystallography and nuclear magnetic resonance (primary techniques) and a broad range of other experimental approaches (secondary techniques). Confident and ambiguous annotations are highlighted separately. DisProt 7.0 presents classified knowledge regarding the experimental characterization and functional annotations of IDPs/IDRs, and is intended to provide an invaluable resource for the research community for a better understanding structural disorder and for developing better computational tools for studying disordered proteins.

INTRODUCTION

Our traditional view of protein structure and function is deeply rooted in the structure–function paradigm which stated that the polypeptide chain of proteins needs to fold into a stable three-dimensional (3D) structure, which is a prerequisite of the functioning of the protein. The extreme explanatory power and success of this model is attested by more than hundred thousand high-resolution structures in the Protein Data Bank (PDB) (1) and many Nobel Prizes awarded for describing structures central to understanding important cell-biological phenomena. It has been suggested almost 20 years ago, however, that many proteins or regions of proteins in various proteomes lack such stable 3D structure, and are rather intrinsically disordered under native, physiological-like conditions (thus named IDPs/IDRs, respectively) (2–4). The recognition of this structural phenomenon brought a radical change in the structure–function paradigm, and critically extended the general appreciation of the role of dynamics in protein function. It has been recognized that structural disorder, which is prevalent in all organisms, plays roles primarily in cellular signaling and regulation (5). Because of that, IDPs/IDRs are often implicated in diseases (6) and represent important drug targets (7).

The structural and functional characterization of disordered proteins represents a special challenge, because they exist as an ensemble of rapidly interconverting conformations. Although they cannot be crystallized and thus cannot be directly characterized by X-ray crystallography, there are a variety of techniques that can report on their highly dynamic structural state at low- or even high spatial and temporal resolution (3). The current best structural description of IDPs/IDRs is by structural ensembles, which can be solved by a combination of experimental and computational approaches and are collected into a dedicated structural database, PED (8).

Studies of the structure–function relationship of disordered proteins have shown that in certain cases their function arises directly from the disordered state (entropic chains), whereas in many other cases their function emanates from molecular recognition accompanied by induced folding to specific binding partners, such as another protein, RNA or DNA molecule (9,10). In these functions, the sensitivity to regulated remodeling of the disordered structural ensemble is an excellent substrate for protein regulation, as exemplified by frequent post-translational modifications (11) and special modes of allosteric regulation (12) involving IDPs/IDRs.

Due to the prevalence and importance of structural disorder, several dedicated databases covering various aspects of IDPs/IDRs have appeared in the past decade. DisProt is the primary repository of disorder-related data on sequence- and functional annotations, focusing on disordered proteins or regions with experimental verification (13,14). Several other databases are based on predictions of disorder, such as D^2P^2, which contains disorder protein predictions by a variety of predictors on 1765 complete proteomes (15), MobiDB, which features three levels of annotations, manually curated, indirect and predicted for all UniProt sequences (over 80 million) (16), and IDEAL, which contains manual annotations of interaction regions undergoing induced folding, sites of post-translational modifications and assignments of structural domains (17). In addition, as already mentioned, PED is the database that gathers structural information on IDPs/IDRs, in the form of structural ensembles (8). The interaction of IDPs/IDRs with their target(s) is most often mediated by short continuous stretches of amino acids such as Molecular Recognition Elements/Features (MoREs/MoRFs) (18) and short/eukaryotic linear motifs (SLiMs/ELMs), which have been collected in the ELM database (19). Less frequently, partner interactions of IDPs/IDRs may also be mediated by intrinsically disordered domains (IDDs), i.e. longer regions that conform to the definition of domains as functional, evolutionary and structural units (20). Although probably still underappreciated, some of these IDDs may be found in the Pfam database of protein families which includes their annotations and underlying multiple sequence alignments (21).

DisProt is central to all IDP-related research efforts, because it collects and presents in a structured way the core experimental evidence reported for structural disorder in proteins. To give a new impetus to the field, we have significantly updated and upgraded it with new features. This new release—DisProt 7.0—contains more than 800 entries of IDPs/IDRs. We have also re-defined and extended functional categories laying the basis for a functional ontology of IDPs, now encompassing 7 major classes and 35 subclasses, all based on published experimental data.

Detection and characterization of IDPs

Technical advances in the field of biophysical and structural biology in the last 50 years have provided the scientific community with an arsenal of techniques to tackle the challenging characterization of IDPs/IDRs (4,22). The various methods differ in their extent of sophistication, and hence in their technical demand, as well as in the nature of the information they provide. Nuclear magnetic resonance (NMR) and X-ray crystallography provide site-specific information, whereas other methods provide more qualitative and global information (e.g. far-UV circular dichroism, size-exclusion chromatography; SEC).

The rise of the field of protein disorder has greatly benefited from structural biology, because structures deposited in the PDB (1) have been instrumental for the development of disorder predictors, often trained on regions of missing electron density. Developments of multidimensional heteronuclear NMR also enabled the structural characterization of disordered proteins of increasing size (23,24). In particular, heteronuclear single quantum coherence (HSQC) experiments are most commonly used to define protein disorder irrespective of whether residue-specific chemical shifts are available or not, as crowded HSQC spectra, characterized by a poor spread of resonances, are typical of IDPs/IDRs. The same feature of low spread of proton resonances is also apparent in one-dimensional proton-based NMR spectra, which offers the obvious advantage of not requiring isotopic labeling. Following assignment of the spectrum, quantitative estimations of disorder can be obtained through various NMR observables, such as chemical shifts, relaxation rates, residual dipolar couplings and resonance intensities in paramagnetic relaxation enhancement experiments. These data enable probing sequence-specific structural information in IDPs/IDRs. A particular strength of NMR is that it can be increasingly applied under truly *in vivo* conditions, in live cells (25). Therefore, these two experimental approaches, X-ray crystallography and multidimensional NMR, are considered as the 'primary techniques' providing evidence for structural disorder on a per residue basis in DisProt.

It should not miss our attention, though, that due to the expenses of isotopic labeling in NMR and the high rate of failure in protein crystallization, it would be unreasonable to only rely on these two approaches to document protein disorder. Therefore, beyond X-ray crystallography and NMR, a plethora of alternative biochemical and biophysical approaches (termed 'secondary techniques') provide orthogonal information on protein disorder in DisProt (4,22). The various approaches are of course not equivalent in terms of reliability, resolution and accuracy and suffer from specific drawbacks and limitations. Structural disorder is often based on far-UV CD spectroscopy, which is overall quite reliable, but does not enable discrimination between ordered and molten globular forms. Near-UV CD, beyond being able to unveil the lack of ordered structure, has the advantage of distinguishing between globular and molten globule forms. Another hallmark of disorder is anomalous sodium dodecyl sulphate-polyacrylamide gel electrophoresis migration, where IDPs have a high apparent molecular mass. IDPs/IDRs also behave anomalously in SEC, light scattering (DLS, MALS), and in small-angle X-ray scattering in that they display hydrodynamic radii (RH) and radii of gyration (Rg) higher than expected, reflecting an extended conformation.

Fluorescence spectroscopy is another common method to assess disorder. Intrinsic fluorescence probing the chemical environment of tryptophan residues provides information about their solvent-accessibility, whereas thermal differential scanning fluorimetry—similar to differential scanning calorimetry—can highlight the lack of a cooperative thermal transition and hence absence of ordered structure. Fluorescence resonance energy transfer between external fluorophores can even generate information on distance distributions and help solve the structural ensemble of the IDP (26). Hyper-sensitivity to proteolysis is also commonly used to map out disordered regions of proteins. Recently, native mass spectrometry exploiting nano-electrospray ionization (27,28) and high-speed atomic force microscopy operating at the single-molecule level (29) have emerged as attractive alternatives to address structural disorder.

As a last statement, it is noteworthy that the higher the number of independent experimental lines supporting disorder, the higher the reliability of the annotation. Furthermore, multi-dimensional information may help realize that structural disorder is not a single homogeneous structural state along an order-disorder binary classification coordinate, it rather represents a continuum of states from the fully ordered to the fully disordered. Similarly, many examples of biological relevant disorder in fragments that are missing from the full length protein have been reported. Furthermore, numerous functional examples of 'conditional disorder', i.e. instances where a disordered region functions by transitions to or from a folded state (30), or when disorder is only observed in a fraction of similar structures (31), lead to ambiguity and clearly points to the need for carrying out complementary experiments. In addition, an extreme case leading to conflicting results is represented by instances where a protein region, predicted to be ordered, is not defined in the electron density in one crystal structure while being ordered in another one (for an example see (32) and DisProt entry DP00133). Do these ambiguous regions represent a new class of disorder that escape detection using the currently available disorder predictors (thus setting the scene for their improvement), or *a contrario* are they the result of static disorder that arises from experimental conditions or domain wobbling? Combining information from a variety of sources may help clarify these cases and also improve meaningful descriptions of IDPs as conformational ensembles (33,34), which may lead to future descriptions of the structure–function relationship of IDPs.

Database structure and implementation

Database records. The technology of DisProt has been updated and is now based on a document-oriented MongoDB database. Stored documents are of two types, 'protein' including general information about the protein and 'disordered region (DR)' including evidence of disorder from literature. Protein information is retrieved from UniProt and includes cleavage sites and chain/peptide boundaries for polyproteins and processed proteins. DisProt is sequence-

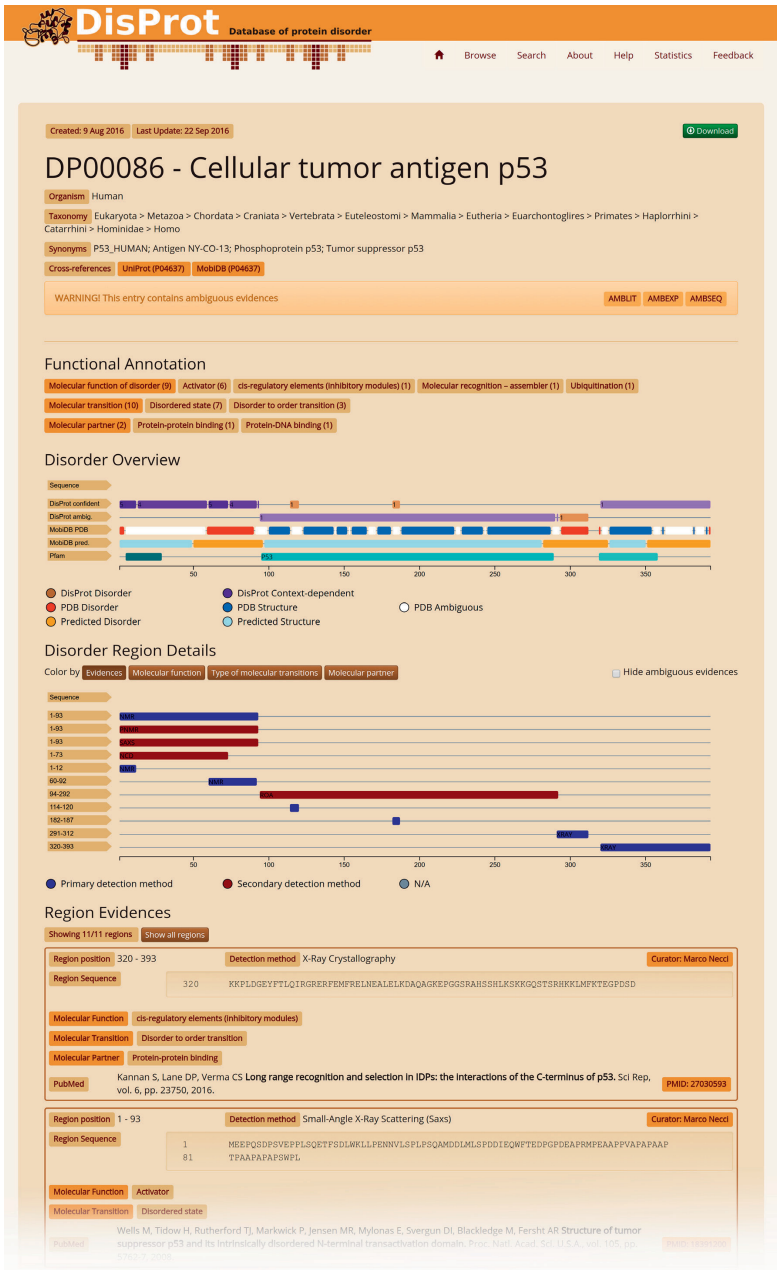


Figure 1. DisProt sample entry, human p53 protein (DP00086). Several experiments have been carried out to characterize the human p53 protein. DisProt reports literature evidence for IDRs. In particular, 11 different IDR evidences (Region Evidences) have been collected from nine different papers by two different curators. Most of these are related to the N-terminus and come from different types of experiments (Disorder Region Details). Disorder regions and the number of DisProt evidences, separated into confident and ambiguous annotations, can be compared with structural information from the Pfam and MobiDB databases in the Disorder Overview. DisProt also provides function annotation of IDRs by reporting molecular function, transition and partner terms (Functional Annotation). A literature reference is provided for each annotated IDR, linked to the relevant PubMed entry.

centric and different isoforms correspond to different entries as in the previous version. Cleaved proteins are merged into a single entry as they are products of the same native sequence. DisProt accession numbers now follow a single format and all previous entries with a '_xxx' suffix were removed. DR records are evidence-centric, i.e. different documents are stored for different experiments even when related to the same region. Forcing a one-to-one paradigm allows to track annotation evidence type and the corresponding literature source unambiguously. DR records also include experimental evidence quality tags for ambiguous annotations. Sometimes experiments are carried out on engineered sequences or fragments which may prove ambiguous to generalize for the entire sequence (AMBSEQ). Moreover, disorder boundaries are occasionally not clear from the literature (AMBLIT) or experiments are performed under extremely non-physiological conditions (AMBEXP). The major improvement from previous versions is the manually curated functional annotation of the regions. Whenever possible, curator-associated functions based on literature evidence are indicated by selecting terms from a new ontology built for describing disorder-related functional modes. If none of the current terms in the new ontology give a proper description of the functional mode, the curator may propose a new term to be added to the ontology. Acceptance of the new term will require approval by the IDP/IDR ontology committee.

Annotation pipeline. The new DisProt data have been generated by a community effort through a web server interface accessible upon registration. The same infrastructure can be used both to create and update entries. Curators provide an annotation through a submission form where all fields are validated on the client-side and a sequence viewer allows the comparison of assigned regions with structure information (Pfam domains, MobiDB disorder). Of note, the name of the curator is clearly visible in the entry to allow proper attribution of credit. The pipeline is fully automatic and can be potentially applied to the entire UniProt database. The DisProt public database is a snapshot of the community annotations.

Entry page. The entry page features four different sections (Figure 1). A protein information table gives the protein name, gene, synonyms, identifiers, taxonomy and 'homologous' entries inferred from sequence similarity. An interactive feature viewer reports DisProt disorder regions separated into confident and ambiguous annotations, colored brown for intrinsically disordered regions and purple for context-dependent regions. Pfam domains along with PDB and predicted disorder derived from MobiDB are also shown. Below, a detailed feature viewer provides different visualization layers to highlight different functional aspects (ontology terms) and the strength of available disorder evidence. Each position in the sequence is colored according to the number and type of evidence. Last but not least, the full curator-generated list of region evidences is reported on the bottom of the page and can be filtered by selecting an element (region) in the feature viewer. Figure 1 shows the current DisProt annotation for the human p53 protein. The combination of DisProt and PDB annotation clearly shows how p53 contains several segments undergoing disorder to order transitions. Evidence for disorder from the literature in the central p53 DNA binding domain, for which many crystal structures are available in the PDB, is ambiguous and highlighted with AMBLIT. Similar conflicts can probably be found in scores of DisProt entries and demonstrate the importance of flagging ambiguous data.

Browsing and searching data. Both browsing and searching functionalities are provided in a single solution from the 'Browse' page. A sortable, customizable and filterable table lists all entries by protein. Alternatively, another table listing all regions is available and accessible through the 'regions' button. Complex queries can be simulated applying different filters to different columns. Specific entries can be selected manually and customized views can be generated by adding or removing columns. Filtered and/or selected data can be downloaded both in text and JSON formats. Alternatively, the 'Search' page allows the user to search for specific words in a free-text form or to search for DisProt entries similar to a query sequence. Output for either search is a provided in a simplified form.

Feedback page. DisProt users are highly encouraged to suggest additional disorder annotations or changes to existing annotations using the 'Feedback' page. This contains a drop-down menu guiding the choice of feedback provided (e.g. website experience, novel annotations) and a message field. For feedback related to data entries, the user is asked to provide either the UniProt or DisProt ID and (where possible) a PubMed reference. All messages are reviewed by the curators and integrated in the database as time permits.

Web technology. The DisProt server is implemented in Node.js (https://nodejs.org) using the REST (Representational State Transfer) architecture. The data can be accessed through the web interface or programmatically exploiting the RESTful functionality. Please refer to the 'Help' section of the website for details on using the DisProt web services. The web interface is built using Angular.js (https://angularjs.org) and Bootstrap (http://getbootstrap.com) frameworks. The feature viewer is implemented on top of the Bio.js library.

Database content: upgrades and updates

Entries in DisProt 7.0 came from three major sources: (i) from the previous version of DisProt (where conflicting cases have been re-annotated), (ii) novel cases identified as PDB entries with long regions of missing electron density and (iii) proteins identified by text-mining in PubMed abstracts for keywords 'intrinsically disordered', 'intrinsically unstructured' and 'structural disorder'. New proteins selected based on disorder content (estimated based on MobiDB data) were prioritized (if appropriate information was available in SwissProt) to concentrate on well-studied and most interesting cases. New proteins were also selected by curators themselves to exploit their specific previous knowledge. All entries from previous versions were re-annotated to remove inconsistencies. One hundred and ninety-eight previous entries were completely removed and 469 modified. Recurring problems being fixed were wrong organism

Table 1. DisProt annotation content

Method/function	Proteins	Regions	Residues
Nuclear magnetic resonance (NMR)	333	592	32 926
X-ray crystallography	326	683	20 742
Circular dichroism (CD) spectroscopy, far-UV	261	352	53 935
Sensitivity to proteolysis	75	95	13 961
Size exclusion/gel filtration chromatography	62	67	12 206
Proton-based NMR	53	69	7723
SDS-PAGE gel, aberrant mobility on	34	34	6326
Other methods	237	273	41 833
Disorder transition	564	1505	151 498
Molecular function	489	1199	106 670
Molecular partner	444	1108	119 665

Distribution of DisProt annotation based on experimental evidence (method) and disorder function (function). As each annotated disorder region corresponds to one piece of experimental evidence, multiple regions can map to the same sequence segment. If a protein is annotated multiple times with the same type of experiment it is counted once. The number of residues is the sum of region lengths.

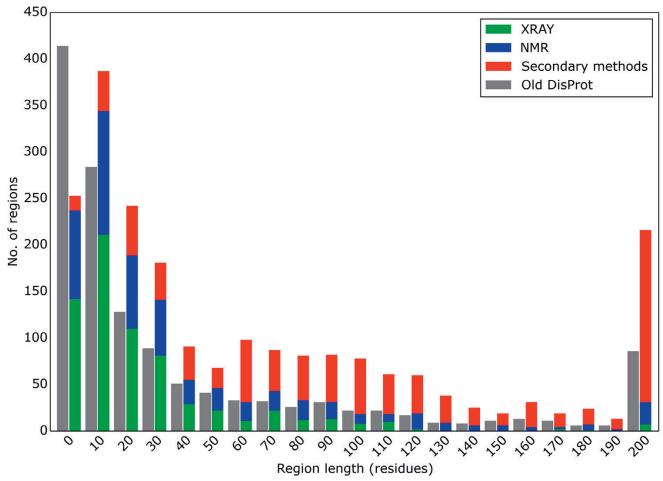


Figure 2. Distribution of disorder segment lengths. Segment lengths are binned in groups of 10 residues, e.g. the column 10 showing lengths between 10 and 19 residues. The current DisProt release is distinguished by experimental technique (X-ray in green, NMR in blue and other methods in red). The previous DisProt release is shown in a single gray bar as it did not have the experimental technique in a machine-readable format.

or isoform assignments, wrong IDR positioning, untracked disorder evidence (e.g. missing explicit literature reference) and weak evidence (e.g. based on very short fragments, please note that the minimal length of an IDR in DisProt 7.0 is 5 residues). Moreover, disorder annotations based on not traceable author/curator statements were discarded. Where necessary, a curator comment now highlights criticisms relative to a given evidence/experiment, e.g. if the experiment has been carried out on an engineered protein. Regions annotated as structured in previous DisProt releases were removed (33 regions). Information related to experiments has been simplified by skipping technical details regarding experimental conditions. However, weak experimental evidence is filtered out by the curator during annotation and tagged with one of three ambiguous labels. Overall, DisProt 7.0 includes 804 entries and 2167 disordered regions, with a total of 92 432 amino acids with clear experimental and functional annotations (Table 1), and the length distribution of disordered regions has significantly changed from the last release of DisProt (Figure 2).

New feature: functional classification

IDPs/IDRs carry out important functions in the cell. The field has settled on the notion that structural disorder represents a continuum of states from fully folded to fully unfolded (random coil-like), and function may come from any of the states and transitions between them. That is, their function may come directly from the disordered state or from molecular recognition and binding to partner molecule(s). We derive our classification from the logic of the gene ontology classification scheme (35), which is based on three structured ontologies ascribing functional terms to gene products (proteins) in terms of their associated biological processes (BP), cellular components (CC) and molecular functions (MF). Apparently, the CC and BP ontologies do not depend on the disordered status of the protein, they

Table 2. Major functional categories of the MFUN ontology of DisProt

MFUN code	Generic functional category	Functional category
MFUN_01	Entropic chain	Flexible linker/spacer
		Entropic bristle
		Entropic clock
		Entropic spring
		Structural mortar
		Self-transport through channel
MFUN_02	Molecular recognition: assembler	Assembler
		Localization (targeting)
		Localization (tethering)
		Prion (self-assembly, polymerization)
		Liquid-liquid phase separation/demixing (self-assembly)
MFUN_03	Molecular recognition: scavenger	Neutralization of toxic molecules
		Metal binding/metal sponge
		Water storage
MFUN_04	Molecular recognition: effector	Inhibitor
		Disassembler
		Activator
		cis-regulatory elements (inhibitory modules)
		DNA bending
		DNA unwinding
MFUN_05	Molecular recognition: display site	Phosphorylation
		Acetylation
		Methylation
		Glycosylation
		Ubiquitination
		Fatty acylation (myristolation and palmitoylation)
		Limited proteolysis
MFUN_06	Molecular recognition: chaperone	Protein detergent/solvate layer
		Space filling
		Entropic exclusion
		Entropy transfer

The functional schemes are an open hierarchy. One goal of sharing information with the community through DisProt is to refine our views of the functional modes of IDPs.

simply reflect the intracellular location of the protein and the BP it participates in, which can be kept without reference for the disordered status (35). The situation is entirely different with MF, which describes the elemental activities of a protein at the molecular level. In this regard, IDPs basically differ from folded proteins, such as enzymes or ligand-binding receptors, because their mode of action and type of function are usually completely different from those of folded proteins. Therefore, we have developed a novel classification scheme that merges and expands previous schemes that suggested thirty (36) and six (9) different categories, to provide classified descriptors for their MFs. Because previous categories (9,36) lacked coherence (for example, they treated structural transitions and interaction partners at the same level), we created a rational scheme that distinguishes these different types of ontologies (cf. Table 2 and ref. (3)).

The three sub-ontologies are as follows: (i) molecular function of disorder (MFUN): describes the type of functional readout of function (such as molecular chaperone); (ii) molecular transition (TRAN) necessary for function (such as disorder-to-order transition); and (iii) molecular partner (PART) that is recognized by the disordered protein (such as protein/RNA/DNA/small molecule). The MFUN ontology is described in detail in Table 1. The TRAN ontology can be further simplified to two IDR states (disorder and transition) to highlight different types of behavior, e.g. in the feature viewer of each DisProt entry.

CONCLUSIONS AND FUTURE WORK

We have presented an updated and completely re-worked version of the DisProt database. It now features state-of-the-art database and web technology, enabling programmatic access of interested parties. The content was expanded by defining a standardized set of experimental techniques and a novel functional ontology of disordered segments. Both allow for a richer description of disorder which may be used for further analyses. The other main improvement in DisProt is a complete re-annotation of existing entries to remove inconsistencies and an expansion of ca. 50% over the previous release, which also resulted in a significant shift in the length coverage of disordered regions in the database. This advance was made possible by a distributed annotation effort coordinated by the COST Action NGP-net (URL: ngp-net.bio.unipd.it) involving a dozen different groups and close to 40 annotators. The longer term maintenance of DisProt is provided by the Italian node of the European bioinformatics infrastructure Elixir. In the future we hope that DisProt can be able to provide disorder annotations for UniProt.

Finally, we hope that the upgrade of DisProt will encourage the scientific community to deposit experimental evidence for disorder within this unique repository, and that this renewed momentum will lead to an increased awareness of the importance of intrinsic disorder in proteins.

FUNDING

COST Action BM1405 NGP-net; ELIXIR-IIB (elixir-italy.org); 'Lendület' Grant from the Hungarian Academy of Sciences [LP2014-16 to Z.D.]; Hungarian Scientific Research Fund [OTKA K 108798 to Z.D.]; AIRC Research Fellowship (to D.P.); Spanish Ministerio de Educación Cultura i Deporte PhD Fellowship (to J.P.); Mexican National Council for Science and Technology (CONACYT) PhD Fellowship [215503 to M.M.-C.]; Grant PortoNeuroDRIve@i3S funded by Norte Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF) (to S.M.-R.); Direction Générale des Armées and Aix-Marseille University PhD Fellowship (to E.Sa.); OTKA Grant [PD-OTKA 108772 to E.Sc.]; French Ministry of National Education, Research and Technology PhD Fellowship (to A.S.); Ministry of Education, Science and Technological Development of the Republic of Serbia [173001, 173049 to N.V., R.D.]; ICREA-Academia Award (to S.V.); Odysseus Grant from Research Foundation Flanders (FWO) [G.0029.12 to P.T.]; AIRC IG Grant [17753 to S.T., in part]; Italian Ministry of Health [GR-2011-02347754 to E.L., S.T.; GR-2011-02346845 to S.T.]; Swedish Research Council Grant [VR-NT 2012-5046 to A.E.]. Computational resources were provided by the Swedish National Infrastructure for Computing (SNIC) at NSC. Funding for open access charge: COST Action BM1405 NGP-net.

Conflict of interest statement. None declared.

REFERENCES

1. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
2. Oldfield,C.J. and Dunker,A.K. (2014) Intrinsically disordered proteins and intrinsically disordered protein regions. *Annu. Rev. Biochem.*, **83**, 553–584.
3. van der Lee,R., Buljan,M., Lang,B., Weatheritt,R.J., Daughdrill,G.W., Dunker,A.K., Fuxreiter,M., Gough,J., Gsponer,J., Jones,D.T. *et al.* (2014) Classification of intrinsically disordered regions and proteins. *Chem. Rev.*, **114**, 6589–6631.
4. Habchi,J., Tompa,P., Longhi,S. and Uversky,V.N. (2014) Introducing protein intrinsic disorder. *Chem. Rev.*, **114**, 6561–6588.
5. Xie,H., Vucetic,S., Iakoucheva,L.M., Oldfield,C.J., Dunker,A.K., Uversky,V.N. and Obradovic,Z. (2007) Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *J. Proteome Res.*, **6**, 1882–1898.
6. Uversky,V.N., Oldfield,C.J. and Dunker,A.K. (2008) Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu. Rev. Biophys.*, **37**, 215–246.
7. Metallo,S.J. (2010) Intrinsically disordered proteins are potential drug targets. *Curr. Opin. Chem. Biol.*, **14**, 481–488.
8. Varadi,M., Kosol,S., Lebrun,P., Valentini,E., Blackledge,M., Dunker,A.K., Felli,I.C., Forman-Kay,J.D., Kriwacki,R.W., Pierattelli,R. *et al.* (2014) pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. *Nucleic Acids Res.*, **42**, D326–D335.
9. Tompa,P. (2005) The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett.*, **579**, 3346–3354.
10. Wright,P.E. and Dyson,H.J. (2009) Linking folding and binding. *Curr. Opin. Struct. Biol.*, **19**, 31–38.
11. Iakoucheva,L.M., Radivojac,P., Brown,C.J., O'Connor,T.R., Sikes,J.G., Obradovic,Z. and Dunker,A.K. (2004) The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.*, **32**, 1037–1049.
12. Tompa,P. (2014) Multisteric regulation by structural disorder in modular signaling proteins: an extension of the concept of allostery. *Chem. Rev.*, **114**, 6715–6732.
13. Vucetic,S., Obradovic,Z., Vacic,V., Radivojac,P., Peng,K., Iakoucheva,L.M., Cortese,M.S., Lawson,J.D., Brown,C.J., Sikes,J.G. *et al.* (2005) DisProt: a database of protein disorder. *Bioinformatics*, **21**, 137–140.
14. Sickmeier,M., Hamilton,J.A., LeGall,T., Vacic,V., Cortese,M.S., Tantos,A., Szabo,B., Tompa,P., Chen,J., Uversky,V.N. *et al.* (2007) DisProt: the Database of Disordered Proteins. *Nucleic Acids Res.*, **35**, D786–D793.
15. Oates,M.E., Romero,P., Ishida,T., Ghalwash,M., Mizianty,M.J., Xue,B., Dosztanyi,Z., Uversky,V.N., Obradovic,Z., Kurgan,L. *et al.* (2013) D(2)P(2): database of disordered protein predictions. *Nucleic Acids Res.*, **41**, D508–D516.
16. Potenza,E., Di Domenico,T., Walsh,I. and Tosatto,S.C. (2015) MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. *Nucleic Acids Res.*, **43**, D315–D320.
17. Fukuchi,S., Sakamoto,S., Nobe,Y., Murakami,S.D., Amemiya,T., Hosoda,K., Koike,R., Hiroaki,H. and Ota,M. (2012) IDEAL: intrinsically disordered proteins with extensive annotations and literature. *Nucleic Acids Res.*, **40**, D507–D511.
18. Mohan,A., Oldfield,C.J., Radivojac,P., Vacic,V., Cortese,M.S., Dunker,A.K. and Uversky,V.N. (2006) Analysis of molecular recognition features (MoRFs). *J. Mol. Biol.*, **362**, 1043–1059.
19. Dinkel,H., Van Roey,K., Michael,S., Kumar,M., Uyar,B., Altenberg,B., Milchevskaya,V., Schneider,M., Kuhn,H., Behrendt,A. *et al.* (2016) ELM 2016–data update and new functionality of the eukaryotic linear motif resource. *Nucleic Acids Res.*, **44**, D294–D300.
20. Tompa,P., Fuxreiter,M., Oldfield,C.J., Simon,I., Dunker,A.K. and Uversky,V.N. (2009) Close encounters of the third kind: disordered domains and the interactions of proteins. *Bioessays*, **31**, 328–335.
21. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
22. Receveur-Brechot,V., Bourhis,J.M., Uversky,V.N., Canard,B. and Longhi,S. (2006) Assessing protein disorder and induced folding. *Proteins*, **62**, 24–45.
23. Kosol,S., Contreras-Martos,S., Cedeno,C. and Tompa,P. (2013) Structural characterization of intrinsically disordered proteins by NMR spectroscopy. *Molecules*, **18**, 10802–10828.
24. Felli,I.C. and Pierattelli,R. (2012) Recent progress in NMR spectroscopy: toward the study of intrinsically disordered proteins of increasing size and complexity. *IUBMB Life*, **64**, 473–481.
25. Theillet,F.X., Binolfi,A., Bekei,B., Martorana,A., Rose,H.M., Stuiver,M., Verzini,S., Lorenz,D., van Rossum,M., Goldfarb,D. *et al.* (2016) Structural disorder of monomeric alpha-synuclein persists in mammalian cells. *Nature*, **530**, 45–50.
26. Schuler,B., Soranno,A., Hofmann,H. and Nettels,D. (2016) Single-molecule FRET spectroscopy and the polymer physics of unfolded and intrinsically disordered Proteins. *Annu. Rev. Biophys.*, **45**, 207–231.
27. Kaltashov,I.A., Bobst,C.E. and Abzalimov,R.R. (2013) Mass spectrometry-based methods to study protein architecture and dynamics. *Protein Sci.*, **22**, 530–544.
28. Borysik,A.J., Kovacs,D., Guharoy,M. and Tompa,P. (2015) Ensemble methods enable a new definition for the solution to gas-phase transfer of intrinsically disordered proteins. *J. Am. Chem. Soc.*, **137**, 13807–13817.
29. Miyagi,A., Tsunaka,Y., Uchihashi,T., Mayanagi,K., Hirose,S., Morikawa,K. and Ando,T. (2008) Visualization of intrinsically disordered regions of proteins by high-speed atomic force microscopy. *Chemphyschem*, **9**, 1859–1866.
30. Jakob,U., Kriwacki,R. and Uversky,V.N. (2014) Conditionally and transiently disordered proteins: awakening cryptic disorder to regulate protein function. *Chem. Rev.*, **114**, 6779–6805.
31. DeForte,S. and Uversky,V.N. (2016) Resolving the ambiguity: making sense of intrinsic disorder when PDB structures disagree. *Protein Sci.*, **25**, 676–688.
32. Blocquel,D., Habchi,J., Durand,E., Sevajol,M., Ferron,F., Erales,J., Papageorgiou,N. and Longhi,S. (2014) Coiled-coil deformations in crystal structures: the measles virus phosphoprotein multimerization

domain as an illustrative example. *Acta Crystallogr. D Biol. Crystallogr.*, **70**, 1589–1603.
33. Sterckx,Y.G., Volkov,A.N., Vranken,W.F., Kragelj,J., Jensen,M.R., Buts,L., Garcia-Pino,A., Jove,T., Van Melderen,L., Blackledge,M. *et al.* (2014) Small-angle X-ray scattering- and nuclear magnetic resonance-derived conformational ensemble of the highly flexible antitoxin PaaA2. *Structure*, **22**, 854–865.
34. Aznauryan,M., Delgado,L., Soranno,A., Nettels,D., Huang,J.R., Labhardt,A.M., Grzesiek,S. and Schuler,B. (2016) Comprehensive structural and dynamical view of an unfolded protein from the combination of single-molecule FRET, NMR, and SAXS. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, E5389–E5398.
35. Consortium, G.O. (2015) Gene ontology consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056.
36. Dunker,A.K., Brown,C.J., Lawson,J.D., Iakoucheva,L.M. and Obradovic,Z. (2002) Intrinsic disorder and protein function. *Biochemistry*, **41**, 6573–6582.

PUBLICATION

III

MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins.

D. Piovesan*, F. Tabaro*, L. Paladin, M. Necci, I. Micetic, C. Camilloni, N. Davey, Z. Dosztányi, B. Mészáros, A. M. Monzon, G. Parisi, E. Schad, P. Sormanni, P. Tompa, M. Vendruscolo, W. F. Vranken and S. C. E. Tosatto

Nucleic acids research 46.(2018), D471–D476
DOI: 10.1093/nar/gkx1071

Published online 10 November 2017 *Nucleic Acids Research, 2018, Vol. 46, Database issue* ***D471–D476***
doi: 10.1093/nar/gkx1071

MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins

Damiano Piovesan[1,†], Francesco Tabaro[1,2,†], Lisanna Paladin[1], Marco Necci[1,3,4], Ivan Mičetić[1], Carlo Camilloni[5], Norman Davey[6,7], Zsuzsanna Dosztányi[8], Bálint Mészáros[8,9], Alexander M. Monzon[10], Gustavo Parisi[10], Eva Schad[9], Pietro Sormanni[11], Peter Tompa[9,12,13], Michele Vendruscolo[11], Wim F. Vranken[12,13,14] and Silvio C.E. Tosatto[1,15,*]

[1]Department of Biomedical Sciences, University of Padua, via U. Bassi 58/b, 35131 Padua, Italy, [2]Institute of Biosciences and Medical Technology, Arvo Ylpön katu 34, 33520 Tampere, Finland, [3]Department of Agricultural Sciences, University of Udine, via Palladio 8, 33100 Udine, Italy, [4]Fondazione Edmund Mach, Via E. Mach 1, 38010 S. Michele all'Adige, Italy, [5]Department of Biosciences, University of Milan, 20133 Milano, Italy, [6]Conway Institute of Biomolecular & Biomedical Research, University College Dublin, Belfield, Dublin 4, Ireland, [7]UCD School of Medicine & Medical Science, University College Dublin, Belfield, Dublin 4, Ireland, [8]MTA-ELTE Lendület Bioinformatics Research Group, Department of Biochemistry, Eötvös Loránd University, 1/c Pázmány Péter sétány, H-1117, Budapest, Hungary, [9]Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, PO Box 7, H-1518 Budapest, Hungary, [10]Structural Bioinformatics Group, Department of Science and Technology, National University of Quilmes, CONICET, Roque Saenz Pena 182, Bernal B1876BXD, Argentina, [11]Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, UK, [12]Structural Biology Brussels, Vrije Universiteit Brussel (VUB), Brussels 1050, Belgium, [13]VIB-VUB Center for Structural Biology, Flanders Institute for Biotechnology (VIB), Brussels 1050, Belgium, [14]Interuniversity Institute of Bioinformatics in Brussels, ULB/VUB, 1050 Brussels, Belgium and [15]CNR Institute of Neuroscience, via U. Bassi 58/b, 35131 Padua, Italy

Received September 22, 2017; Revised October 13, 2017; Editorial Decision October 16, 2017; Accepted October 19, 2017

ABSTRACT

The MobiDB (URL: mobidb.bio.unipd.it) database of protein disorder and mobility annotations has been significantly updated and upgraded since its last major renewal in 2014. Several curated datasets for intrinsic disorder and folding upon binding have been integrated from specialized databases. The indirect evidence has also been expanded to better capture information available in the PDB, such as high temperature residues in X-ray structures and overall conformational diversity. Novel nuclear magnetic resonance chemical shift data provides an additional experimental information layer on conformational dynamics. Predictions have been expanded to provide new types of annotation on backbone rigidity, secondary structure preference and disordered binding regions. MobiDB 3.0 contains information for the complete UniProt protein set and synchronization has been improved by covering all UniParc sequences. An advanced search function allows the creation of a wide array of custom-made datasets for download and further analysis. A large amount of information and cross-links to more specialized databases are intended to make MobiDB the central resource for the scientific community working on protein intrinsic disorder and mobility.

INTRODUCTION

The protein structure-function paradigm is a cornerstone of molecular biology, offering a mechanistic understanding of processes ranging from enzyme catalysis, signal transduction to molecular recognition and allosteric regulation. Underlying this paradigm is the assumption that proteins become functional by assuming a well-defined structure, typically described by the coordinates of all its atoms. A solid foundation of this view is provided by the 130 000 structures of proteins and complexes in the Protein Data Bank, PDB (1). However, it is increasingly recognized that many proteins do not obey this rule. Intrinsically disordered proteins (IDPs) or regions (IDRs) are devoid of order in their native unbound state (2–4). Intrinsic disorder is prevalent

*To whom correspondence should be addressed. Tel: +39 49 827 6269; Fax: +39 49 827 6363; Email: silvio.tosatto@unipd.it
†These authors contributed equally to the paper as first authors.

in the human proteome (5), appears to play important signaling and regulatory roles (2) and is frequently involved in disease (6). The discovery of intrinsic disorder and its prevalence and functional importance is transforming the field of molecular biology. As intrinsic disorder is emerging as a general phenomenon, databases are collecting and presenting disorder related data in a systematic manner. MobiDB has been a major contributor by providing consensus predictions and functional annotations for all UniProt proteins, driving the field ahead (7,8). The MobiDB upgrade we present in this paper is essential for several reasons.

There is a rapid advance in the functional understanding of intrinsic disorder. The functional classification of IDPs/IDRs is becoming ever more elaborate, with several newly recognized functional mechanisms (9). For example, the central role of intrinsic disorder in the formation of membraneless organelles, such as nucleoli and stress granules, by liquid-liquid phase separation has been characterized recently (10–13). A wide range of experimental observations on the structure-function relationship of IDPs/IDRs is furthering our understanding of disordered states and of the manners in which they function (14–16). These developments have also played a central role in the recent update of the DisProt database (17), the central repository of experimentally characterized IDPs and IDRs. The re-curated version of this database contains experimental observations of disorder for more than 800 protein entries and a renewed functional ontology schema. The experimental evidence on which it rests has also been significantly augmented to include a broad range of biophysical techniques. DisProt is the basis for most developments in disorder predictors (18,19), and its recent update is a major motivation for a new version of MobiDB.

Additional developments in the field make this release timely. A major source of intrinsic disorder is the identification of residues with missing atomic coordinates in the PDB, which can now be augmented by cryo-electron microscopy (cryo-EM) data. This is having a tremendous impact on structural biology (20,21). Structural descriptions of IDPs and IDRs under physiological conditions have also greatly advanced and are starting to appear in dedicated databases such as IDEAL (22), DIBS (23) and MFIB (24). IDPs and IDRs can perform key roles in molecular recognition by folding upon binding of short linear motifs (SLiMs) covered in the ELM database (25). Generally, the full functional characterization of IDPs and IDRs requires the description not just of their free (disordered) states (26,27), but also of their residual dynamics in the bound states (28). Fuzzy (disordered) complexes can be found in FuzDB (29) and structural ensembles describing the free form (30) in the protein ensemble database (PED (31)). Techniques such as in-cell Nuclear magnetic resonance (NMR) spectroscopy (32,33) and single-molecule fluorescence (34) will soon help study these structures in the physiological state. In reflection of all these developments, we are now launching a significantly updated version of our database, MobiDB 3.0. The new version incorporates additional curated data from specialized databases. Novel annotation features include disorder derived from publicly available NMR chemical shift data (35) and an extended list of predictors. Database

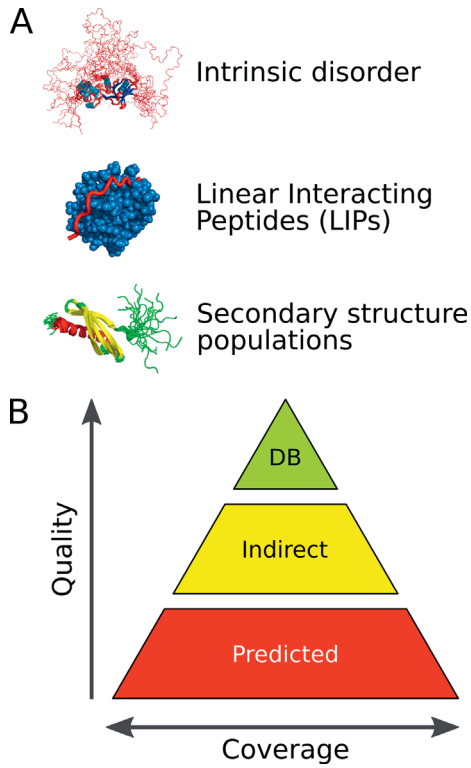


Figure 1. Overview of different annotation data types (**A**) and levels of accuracy (**B**) in MobiDB 3.0.

searches are facilitated by an improved search algorithm, pre-calculated data and new sections in the database.

DATABASE DESCRIPTION

MobiDB 3.0 is intended to be a central resource for large-scale intrinsic disorder sequence annotation. This new version is organized by both type of disorder annotation and quality of disorder evidence (Figure 1). Disorder information is grouped in three different sections: disorder, linear interacting peptides (LIPs) and secondary structure populations. The latter represents the conformational heterogeneity of IDPs and IDRs as the ability to populate different secondary structure populations in solution. LIPs are structure fragments that interact with other molecules preserving an elongated structure or folding upon binding. The data in MobiDB is organized hierarchically. The top tier is formed by manually curated data from external databases and represents the highest quality annotations. Annotations derived from experimental data such as X-ray and NMR chemical shifts are indirect but far more abundant. At the bottom, predictions provide disorder annotation at lower confidence than experimental evidence. The main disorder definition in MobiDB is provided by a consensus combin-

Table 1. Overview of databases integrated into MobiDB 3.0

Database	Type	Comment	URL
UniProt	Curated	Disorder	http://www.uniprot.org/
DisProt	Curated	Disorder	http://www.disprot.org/
FuzDB	Curated	Disorder	http://protdyn-database.org/
ELM	Curated	LIPs	http://elm.eu.org/
MFIB	Curated	LIPs	http://mfib.enzim.ttk.mta.hu/
DIBS	Curated	LIPs	http://dibs.enzim.ttk.mta.hu/
IDEAL	Curated	LIPs	http://www.ideal.force.cs.is.nagoya-u.ac.jp/IDEAL/
Gene3D	Curated/Prediction	Structure	http://gene3d.biochem.ucl.ac.uk/
Pfam	Curated/Prediction	Domains/Families	http://pfam.xfam.org/
CoDNaS	Indirect	Conformational diversity	http://ufq.unq.edu.ar/codnas/

ing all available sources prioritizing curated and indirect evidences over predictions in analogy to the previous version (8). In the following, we will describe the main recent improvements since the previous release. The database schema, web interface and server have been completely redesigned and the underlying technology updated. The feature viewer showing sequence annotations is now fully dynamic and allows the generation of high quality images for publications with a click. Where available, MobiDB annotation is projected directly onto the structure and shown in a new 3D viewer. The look and feel and organization of the page and loading latency were also improved.

New curated data

MobiDB 3.0 includes different sources of manually curated disorder annotations (Table 1). These annotations fall into two categories: disorder and LIPs. LIPs are binding regions presumed or demonstrated to be intrinsically disordered that fold upon binding. These come under different names such as SLiMs or MoREs (molecular recognition elements) in the literature. The IDEAL database calls them 'protean' segments (ProS) (22). MobiDB includes both 'verified' and 'possible' ProS from IDEAL, where verified means disorder has been experimentally observed in the isolated molecule. The Database of Disordered Binding Sites (DIBS, (23)) collects cases where a disordered region folds upon binding with a globular domain and the Mutual Folding Induced by Binding (MFIB, (24)) database includes disordered regions that fold upon binding with another disordered region. ELM (25) provides SLiM annotations involved in binding and post-translational modifications. General disorder annotation, i.e. without any knowledge about transition driven by interactions, is collected from UniProtKB (36), DisProt (17) and FuzDB (29). UniProtKB provides manually curated disorder annotations under the region field in the features section. FuzDB collects cases of fuzzy complexes, where conformational diversity has a functional role in the regulation and formation of protein complexes or higher-order assemblies. DisProt has been recently revamped and MobiDB now propagates DisProt disordered regions by homology transfer. Regions homologous to experimentally characterized IDRs are mapped across homologs obtained from GeneTree alignments (37). Regions with identity and similarity >80% and an alignment of at least 10 residues are retained as homologous IDRs. Gene3D (38) contributes complementary order annotation to the MobiDB consensus calculation, while Pfam (39) is used to highlight protein domains. Lastly MobiDB also maps CoDNaS information to highlight conformation diversity in globular regions. CoDNaS measures structural differences among conformers of the same protein (40).

New indirect annotations

Previous releases of MobiDB provided indirect annotations from the PDB through missing residues in X-ray structures and mobile regions from NMR ensembles as calculated with the Mobi software (41). In the current release, this annotation has been complemented with additional indirect information from experimental data in the PDB and chemical shifts from the Biological Magnetic Resonance Data Bank (BMRB) (35). The new Mobi 2.0 software (42) is used to extract LIPs and disorder information from PDB files. Disorder is encoded by three different parameters: high-temperature, missing and mobile residues. High-temperature residues are detected from B-factor regions for X-ray and cryo-EM structures using a threshold proportional to the resolution of the structure. Missing residues are available for all experimental types and obtained comparing the experimental sequence (i.e. PDB SEQRES entries) with the observed residues in the structure (i.e. PDB ATOM entries). A mobility estimate is provided for NMR structures by comparing C_α displacement and local conformations in different aligned models (41). LIPs are identified by comparing intra- versus inter-chain contacts calculated using RING (43). The closest atoms between two residues are used to establish a contact which is then distinguished by chemical type (e.g. hydrogen bond, salt bridge, $\pi-\pi$ stack). LIPs are identified as any region where the number of inter-chain contacts is at least two times the number of intra-chain contacts (42).

MobiDB 3.0 better exploits the power of NMR spectroscopy to probe the structural properties of proteins in solution, as well as their dynamics on a wide range of timescales (44). Chemical shifts quantify structural fluctuations of proteins up to the millisecond timescale and are relatively easy to measure. Using chemical shifts to obtain information about the statistical populations of different structural motifs allows for a more comprehensive structural description of proteins in solution than static structures or binary definitions such as 'ordered' and 'disordered' (44). MobiDB 3.0 uses chemical shift data from BMRB directly as reported without applying chemical shift re-referencing methods. The software packages δ2D (45) and Random Coil Index (RCI) (46) are used to calculate

Table 2. Overview of tools used into MobiDB 3.0

Tool	Type	Description
Mobi 2.0	Indirect	Missing, high-temperature and mobile residues from PDB structures
RING 2.0	Indirect	Residue interactions from PDB structures, used to define LIPs
RCI	Indirect	Random coil index from BMRB chemical shifts
δ2D	Indirect	Secondary structure populations from BMRB chemical shifts
DynaMine	Prediction	Random coil index
FeSS	Prediction	Secondary structure prediction component of FELLS
MobiDB-lite	Prediction	Long disorder based on consensus
DisEMBL	Prediction	Disorder. Versions: 465, Hot-loops
ESpritz	Prediction	Disorder. Versions: DisProt, NMR, X-ray
IUPred	Prediction	Disorder. Versions: Short, Long
VSL2b	Prediction	Disorder
GlobPlot	Prediction	Globular regions, used as opposite of disorder
SEG	Prediction	Low complexity
Pfilt	Prediction	Low complexity

two-dimensional ensembles in terms of secondary structure populations (44) and backbone flexibility. Secondary structure populations are calculated only for residues with at least three atom types with measured chemical shifts, as using fewer chemical shifts results in less accurate mappings of the populations (45). MobiDB 3.0 reports the experimental conditions at which the chemical shifts were measured as the structural properties of some proteins can change drastically between different conditions (e.g. binding partners, lipids, pH) and these can help elucidate protein function (44). When an entry in MobiDB is associated to multiple chemical shifts, an overview of the predominant secondary structure conformation is provided in a consensus track. This can be expanded in the feature viewer to show experimental conditions such as pH, temperature, binding partners, molecular state, sample information and the title of the corresponding BMRB entries.

New predictors

MobiDB 3.0 includes the same set of disorder predictors used in the previous release: ESpritz (47), IUpred (48), DisEMBL (49) and VSL2b (50). Consensus generation is handled by MobiDB-lite (51), which uses a stronger majority threshold and enforces at least 20 consecutive disordered residues to provide highly specific predictions. This is completed by a continuous representation of the fraction of methods predicting disorder for each residue. DynaMine (52), Anchor (53) and FeSS (54) are now also part of the annotation pipeline. DynaMine (52) predicts backbone flexibility where 1.0 means complete order (stable conformation, i.e. rigid) and 0 means fully random bond vector movement (highly dynamic, i.e. flexible). Anchor predicts binding regions located in disordered proteins, providing LIP annotations for all proteins in the database. FeSS is a component of the FELLS method (54) providing three-state (helix, sheet, coil) secondary structure propensity. FeSS prediction confidence can be interpreted similarly to the dynamic behavior measured by δ2D in chemical shifts, i.e. a propensity to remain in a given state of secondary structure. The complete list of tools is available in Table 2.

The MobiDB-lite version used in MobiDB 3.0 has been extended to provide a structural characterization of the disorder regions that can help interpret their functional role. It distinguishes different types of disordered regions by measuring the fraction of charged residues and net charge according to a previous classification (55). The different types are: positive polyelectrolites (D_PPE), negative polyelectrolites (D_NPE), polyampholites (D_PA) and weak polyampholites (D_WC). A statistical analysis of the different disorder flavours was already performed on the MobiDB 2.0 data (8).

Usage and annotated data

MobiDB now contains all sequences from UniParc, the most comprehensive non-redundant set of protein sequences. Entries are identified also by UniProtKB (36) accession numbers and can be retrieved by organism, taxonomy and other identifiers provided by UniProtKB. Prediction results are combined with indirect disorder evidences derived from PDB data (using Mobi 2) and data extracted from manually curated third party databases. MobiDB annotations are used by DisProt (17) curators to guide the annotation of disorder regions. MobiDB data is made available to the public via a web interface allowing extensive search functionalities and RESTful services for programmatic access. MobiDB 3.0 includes a pre-calculated consensus for all entries allowing real-time statistics and download of entire datasets in different formats directly from the web interface. The new database schema makes it possible to perform complex search queries and to generate custom datasets, for example retrieving all entries with manually curated annotations. The MobiDB update has been automatized and is scheduled every three months due to the high computational cost of generating predictions for new sequences.

DISCUSSION

MobiDB 3.0 improves on previous releases by adding descriptions of conformational diversity and disorder-related functions, both in terms of experimental data and predictions. A particular field where it may have a significant impact is the establishment of a long-awaited disorder sequence-function relationship schema. The most reliable proxy to this goal is to assess the function of a protein by homology transfer, i.e. transferring functional annotation based on sequence similarity. Aligning IDR sequences is complicated by their high evolutionary variabil-

ity and often limits evolutionary analysis (56,57). New functional terms introduced in the DisProt update (17), represent non-canonical functions probably only characteristic of IDPs which are not incorporated in functional classification schemes such as GO (58). A large-scale analysis of IDP functional annotations will be necessary to find adequate boundaries for transferring IDP functions by homology. As sufficient data is now available in MobiDB 3.0, we expect a rapid advance in the field of sequence-function correlations of IDPs.

For proteins with sufficient NMR data, MobiDB now features quantitative annotations incorporating structure and equilibrium dynamics in a unified framework. These large-scale quantitative annotations will help understand the biological role of order and disorder, and serve as a basis to construct predictive models. As NMR measurements of proteins in their native complex environments, such as inside living cells, are becoming more common (59), we will be able to address fundamental biological questions with greater physiological relevance (60).

MobiDB is widely used by scientific community and by third party services such as DisProt (17) and ProViz (61). It has recently joined the InterPro consortium to provide disorder annotation alongside protein domains and families (62). MobiDB is becoming a thematic hub for IDPs inside the European sustainable bioinformatics infrastructure (ELIXIR) and we encourage contributions of novel predictors and datasets. Future work will focus on including IDP annotations into core data resources such as UniProt.

ACKNOWLEDGEMENTS

We acknowledge ELIXIR-IIB (elixir-italy.org), the Italian Node of the European ELIXIR infrastructure (elixir-europe.org), for supporting the development and maintenance of MobiDB.

FUNDING

COST Action [BM1405 NGP-net]; FIRC Research Fellowship [16621 to D.P.]; Hungarian Academy of Sciences ‘Lendület’ Grant [LP201418/2016 to Z.D.]; Hungarian Scientific Research Fund [OTKA K 108798 to Z.D.]; Hungarian Academy of Sciences Postdoctoral Fellowship [to B.M.]; Agencia de Ciencia y Tecnología [PICT-2014–3430 to G.P.]; Universidad Nacional de Quilmes [1402/15]; OTKA Grant [PD-OTKA 108772 to E.S.]; Research Foundation Flanders (FWO) Odysseus Grant [G.0029.12 to P.T.]; FWO project [G032816N to W.F.V.]; AIRC IG Grant [17753 to S.T., in part]. Funding for open access charge: COST Action [BM1405].

Conflict of interest statement. None declared.

REFERENCES

1. Burley,S.K., Berman,H.M., Kleywegt,G.J., Markley,J.L., Nakamura,H. and Velankar,S. (2017) Protein Data Bank (PDB): the single global macromolecular structure archive. *Methods Mol. Biol. Clifton NJ*, **1607**, 627–641.
2. Wright,P.E. and Dyson,H.J. (2015) Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.*, **16**, 18–29.
3. Habchi,J., Tompa,P., Longhi,S. and Uversky,V.N. (2014) Introducing protein intrinsic disorder. *Chem. Rev.*, **114**, 6561–6588.
4. Tompa,P. (2012) Intrinsically disordered proteins: a 10-year recap. *Trends Biochem. Sci.*, **37**, 509–516.
5. Pancsa,R. and Tompa,P. (2012) Structural disorder in eukaryotes. *PLoS One*, **7**, e34687.
6. Uversky,V.N., Oldfield,C.J. and Dunker,A.K. (2008) Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu. Rev. Biophys.*, **37**, 215–246.
7. Di Domenico,T., Walsh,I., Martin,A.J.M. and Tosatto,S.C.E. (2012) MobiDB: a comprehensive database of intrinsic protein disorder annotations. *Bioinformatics*, **28**, 2080–2081.
8. Potenza,E., Di Domenico,T., Walsh,I. and Tosatto,S.C.E. (2015) MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. *Nucleic Acids Res.*, **43**, D315–D320.
9. van der Lee,R., Buljan,M., Lang,B., Weatheritt,R.J., Daughdrill,G.W., Dunker,A.K., Fuxreiter,M., Gough,J., Gsponer,J., Jones,D.T. *et al.* (2014) Classification of intrinsically disordered regions and proteins. *Chem. Rev.*, **114**, 6589–6631.
10. Hyman,A.A., Weber,C.A. and Jülicher,F. (2014) Liquid-liquid phase separation in biology. *Annu. Rev. Cell Dev. Biol.*, **30**, 39–58.
11. Shorter,J. (2016) Membraneless organelles: phasing in and out. *Nat. Chem.*, **8**, 528–530.
12. Toretsky,J.A. and Wright,P.E. (2014) Assemblages: functional units formed by cellular phase separation. *J. Cell Biol.*, **206**, 579–588.
13. Brangwynne,C.P., Tompa,P. and Pappu,R.V. (2015) Polymer physics of intracellular phase transitions. *Nat. Phys.*, **11**, 899–904.
14. Berlow,R.B., Dyson,H.J. and Wright,P.E. (2017) Hypersensitive termination of the hypoxic response by a disordered protein switch. *Nature*, **543**, 447–451.
15. Mylona,A., Theillet,F.-X., Foster,C., Cheng,T.M., Miralles,F., Bates,P.A., Selenko,P. and Treisman,R. (2016) Opposing effects of Elk-1 multisite phosphorylation shape its response to ERK activation. *Science*, **354**, 233–237.
16. Bah,A., Vernon,R.M., Siddiqui,Z., Krzeminski,M., Muhandiram,R., Zhao,C., Sonenberg,N., Kay,L.E. and Forman-Kay,J.D. (2015) Folding of an intrinsically disordered protein by phosphorylation as a regulatory switch. *Nature*, **519**, 106–109.
17. Piovesan,D., Tabaro,F., Mičetić,I., Necci,M., Quaglia,F., Oldfield,C.J., Aspromonte,M.C., Davey,N.E., Davidović,R., Dosztányi,Z. *et al.* (2017) DisProt 7.0: a major update of the database of disordered proteins. *Nucleic Acids Res.*, **45**, D1123–D1124.
18. Walsh,I., Giollo,M., Di Domenico,T., Ferrari,C., Zimmermann,O. and Tosatto,S.C.E. (2015) Comprehensive large-scale assessment of intrinsic protein disorder. *Bioinformatics*, **31**, 201–208.
19. Necci,M., Piovesan,D., Dosztanyi,Z., Tompa,P. and Tosatto,S.C.E. (2017) A comprehensive assessment of long intrinsic protein disorder from the DisProt database. *Bioinformatics*, doi:10.1093/bioinformatics/btx590.
20. Callaway,E. (2015) The revolution will not be crystallized: a new method sweeps through structural biology. *Nature*, **525**, 172–174.
21. Cheng,Y. (2015) Single-Particle Cryo-EM at Crystallographic Resolution. *Cell*, **161**, 450–457.
22. Fukuchi,S., Amemiya,T., Sakamoto,S., Nobe,Y., Hosoda,K., Kado,Y., Murakami,S.D., Koike,R., Hiroaki,H. and Ota,M. (2014) IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners. *Nucleic Acids Res.*, **42**, D320–D325.
23. Schad,E., Fichó,E., Pancsa,R., Simon,I., Dosztányi,Z. and Mészáros,B. (2017) DIBS: a repository of disordered binding sites mediating interactions with ordered proteins. *Bioinformatics*, doi:10.1093/bioinformatics/btx640.
24. Fichó,E., Reményi,I., Simon,I. and Mészáros,B. (2017) MFIB: a repository of protein complexes with mutual folding induced by binding. *Bioinformatics*, doi:10.1093/bioinformatics/btx486.
25. Dinkel,H., Van Roey,K., Michael,S., Kumar,M., Uyar,B., Altenberg,B., Milchevskaya,V., Schneider,M., Kühn,H., Behrendt,A. *et al.* (2016) ELM 2016–data update and new functionality of the eukaryotic linear motif resource. *Nucleic Acids Res.*, **44**, D294–D300.
26. Arai,M., Sugase,K., Dyson,H.J. and Wright,P.E. (2015) Conformational propensities of intrinsically disordered proteins influence the mechanism of binding and folding. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 9614–9619.
27. Borcherds,W., Theillet,F.-X., Katzer,A., Finzel,A., Mishall,K.M., Powell,A.T., Wu,H., Manieri,W., Dieterich,C., Selenko,P. *et al.* (2014)

Disorder and residual helicity alter p53-Mdm2 binding affinity and signaling in cells. *Nat. Chem. Biol.*, **10**, 1000–1002.
28. Tompa,P. and Fuxreiter,M. (2008) Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem. Sci.*, **33**, 2–8.
29. Miskei,M., Antal,C. and Fuxreiter,M. (2017) FuzDB: database of fuzzy complexes, a tool to develop stochastic structure-function relationships for protein complexes and higher-order assemblies. *Nucleic Acids Res.*, **45**, D228–D235.
30. Tompa,P. and Varadi,M. (2014) Predicting the predictive power of IDP ensembles. *Structure*, **22**, 177–178.
31. Varadi,M., Kosol,S., Lebrun,P., Valentini,E., Blackledge,M., Dunker,A.K., Felli,I.C., Forman-Kay,J.D., Kriwacki,R.W., Pierattelli,R. *et al.* (2014) pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. *Nucleic Acids Res.*, **42**, D326–D335.
32. Theillet,F.-X., Binolfi,A., Bekei,B., Martorana,A., Rose,H.M., Stuiver,M., Verzini,S., Lorenz,D., van Rossum,M., Goldfarb,D. *et al.* (2016) Structural disorder of monomeric α-synuclein persists in mammalian cells. *Nature*, **530**, 45–50.
33. Felli,I.C., Gonnelli,L. and Pierattelli,R. (2014) In-cell ^{13}C NMR spectroscopy for the study of intrinsically disordered proteins. *Nat. Protoc.*, **9**, 2005–2016.
34. Sakon,J.J. and Weninger,K.R. (2010) Detecting the conformation of individual proteins in live cells. *Nat. Methods*, **7**, 203–205.
35. Ulrich,E.L., Akutsu,H., Doreleijers,J.F., Harano,Y., Ioannidis,Y.E., Lin,J., Livny,M., Mading,S., Maziuk,D., Miller,Z. *et al.* (2008) BioMagResBank. *Nucleic Acids Res.*, **36**, D402–D408.
36. Consortium,The UniProt (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.
37. Vilella,A.J., Severin,J., Ureta-Vidal,A., Heng,L., Durbin,R. and Birney,E. (2009) EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, **19**, 327–335.
38. Lees,J., Yeats,C., Perkins,J., Sillitoe,I., Rentzsch,R., Dessailly,B.H. and Orengo,C. (2012) Gene3D: a domain-based resource for comparative genomics, functional annotation and protein network analysis. *Nucleic Acids Res.*, **40**, D465–D471.
39. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
40. Monzon,A.M., Rohr,C.O., Fornasari,M.S. and Parisi,G. (2016) CoDNaS 2.0: a comprehensive database of protein conformational diversity in the native state. *Database (Oxford)*, **2016**, baw038.
41. Martin,A.J.M., Walsh,I. and Tosatto,S.C.E. (2010) MOBI: a web server to define and visualize structural mobility in NMR protein ensembles. *Bioinformatics*, **26**, 2916–2917.
42. Piovesan,D. and Tosatto,S.C.E. (2017) Mobi 2.0: an improved method to define intrinsic disorder, mobility and linear binding regions in protein structures. *Bioinformatics*, doi:10.1093/bioinformatics/btx592.
43. Piovesan,D., Minervini,G. and Tosatto,S.C.E. (2016) The RING 2.0 web server for high quality residue interaction networks. *Nucleic Acids Res.*, **44**, W367–W374.
44. Sormanni,P., Piovesan,D., Heller,G.T., Bonomi,M., Kukic,P., Camilloni,C., Fuxreiter,M., Dosztanyi,Z., Pappu,R.V., Babu,M.M. *et al.* (2017) Simultaneous quantification of protein order and disorder. *Nat. Chem. Biol.*, **13**, 339–342.
45. Camilloni,C., De Simone,A., Vranken,W.F. and Vendruscolo,M. (2012) Determination of secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts. *Biochemistry*, **51**, 2224–2231.
46. Berjanskii,M.V. and Wishart,D.S. (2005) A simple method to predict protein flexibility using secondary chemical shifts. *J. Am. Chem. Soc.*, **127**, 14970–14971.
47. Walsh,I., Martin,A.J.M., Di Domenico,T. and Tosatto,S.C.E. (2012) ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics*, **28**, 503–509.
48. Dosztányi,Z., Csizmok,V., Tompa,P. and Simon,I. (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **21**, 3433–3434.
49. Linding,R., Jensen,L.J., Diella,F., Bork,P., Gibson,T.J. and Russell,R.B. (2003) Protein disorder prediction: implications for structural proteomics. *Structure*, **11**, 1453–1459.
50. Peng,K., Radivojac,P., Vucetic,S., Dunker,A.K. and Obradovic,Z. (2006) Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*, **7**, 208.
51. Necci,M., Piovesan,D., Dosztányi,Z. and Tosatto,S.C.E. (2017) MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins. *Bioinformatics*, **33**, 1402–1404.
52. Cilia,E., Pancsa,R., Tompa,P., Lenaerts,T. and Vranken,W.F. (2013) From protein sequence to dynamics and disorder with DynaMine. *Nat. Commun.*, **4**, 2741.
53. Mészáros,B., Simon,I. and Dosztányi,Z. (2009) Prediction of protein binding regions in disordered proteins. *PLoS Comput. Biol.*, **5**, e1000376.
54. Piovesan,D., Walsh,I., Minervini,G. and Tosatto,S.C.E. (2017) FELLS: fast estimator of latent local structure. *Bioinformatics*, **33**, 1889–1891.
55. Das,R.K. and Pappu,R.V. (2013) Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 13392–13397.
56. Brown,C.J., Johnson,A.K., Dunker,A.K. and Daughdrill,G.W. (2011) Evolution and disorder. *Curr. Opin. Struct. Biol.*, **21**, 441–446.
57. Csizmok,V., Felli,I.C., Tompa,P., Banci,L. and Bertini,I. (2008) Structural and dynamic characterization of intrinsically disordered human securin by NMR spectroscopy. *J. Am. Chem. Soc.*, **130**, 16873–16879.
58. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
59. Plitzko,J.M., Schuler,B. and Selenko,P. (2017) Structural Biology outside the box-inside the cell. *Curr. Opin. Struct. Biol.*, **46**, 110–121.
60. Gierasch,L.M. and Gershenson,A. (2009) Post-reductionist protein science, or putting Humpty Dumpty back together again. *Nat. Chem. Biol.*, **5**, 774–777.
61. Jehl,P., Manguy,J., Shields,D.C., Higgins,D.G. and Davey,N.E. (2016) ProViz-a web-based visualization tool to investigate the functional and evolutionary features of protein sequences. *Nucleic Acids Res.*, **44**, W11–W15.
62. Finn,R.D., Attwood,T.K., Babbitt,P.C., Bateman,A., Bork,P., Bridge,A.J., Chang,H.-Y., Dosztányi,Z., El-Gebali,S., Fraser,M. *et al.* (2017) InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res.*, **45**, D190–D199.