Molecular Biology of the Cell
Vol. 16, 1268–1281, March 2005

Growth Factor Receptor Binding Protein 2-mediated Recruitment of the RING Domain of Cbl to the Epidermal Growth Factor Receptor Is Essential and Sufficient to Support Receptor Endocytosis[V]

Fangtian Huang and Alexander Sorkin

Department of Pharmacology, University of Colorado Health Sciences Center, Aurora, CO 80045

Submitted September 23, 2004; Revised December 14, 2004; Accepted December 22, 2004
Monitoring Editor: Sandra Schmid

Knockdown of growth factor receptor binding protein 2 (Grb2) by RNA interference strongly inhibits clathrin-mediated endocytosis of the epidermal growth factor receptor (EGFR). To gain insights into the function of Grb2 in EGFR endocytosis, we have generated cell lines in which endogenous Grb2 was replaced by yellow fluorescent protein (YFP)-tagged Grb2 expressed at the physiological level. In these cells, Grb2-YFP fully reversed the inhibitory effect of Grb2 knockdown on EGFR endocytosis and, moreover, trafficked together with EGFR during endocytosis. Overexpression of Grb2-binding protein c-Cbl did not restore endocytosis in Grb2-depleted cells. However, EGFR endocytosis was rescued in Grb2-depleted cells by chimeric proteins consisting of the Src homology (SH) 2 domain of Grb2 fused to c-Cbl. The "knockdown and rescue" analysis revealed that the expression of Cbl-Grb2/SH2 fusions containing RING finger domain of Cbl restores normal ubiquitylation and internalization of the EGFR in the absence of Grb2, consistent with the important role of the RING domain in EGFR endocytosis. In contrast, the carboxy-terminal domain of Cbl, when attached to Grb2 SH2 domain, had 4 times smaller endocytosis-rescue effect compared with the RING-containing chimeras. Together, the data suggest that the interaction of Cbl carboxy terminus with CIN85 has a minor and a redundant role in EGFR internalization. We concluded that Grb2-mediated recruitment of the functional RING domain of Cbl to the EGFR is essential and sufficient to support receptor endocytosis.

INTRODUCTION

Activation of the epidermal growth factor (EGF) receptor (EGFR) triggers multiple signal transduction events and accelerates endocytosis of EGFR through clathrin-coated pits. Clathrin-mediated endocytosis and subsequent targeting of internalized EGF–receptor complexes to lysosomes result in down-regulation of EGFR protein levels and attenuation of receptor signaling. Endocytosis also controls subcellular localization of activated receptors and their signaling complexes. However, the molecular mechanisms of EGFR endocytosis are not fully understood.

Growth factor receptor binding protein 2 (Grb2) is an essential component of EGFR signaling to Ras (Li *et al*., 1993; Rozakis-Adcock *et al*., 1993). The Src homology (SH) 2 domain of Grb2 can bind directly to phosphotyrosines 1068 and 1086 of the activated EGFR or indirectly through the tyrosine-phosphorylated adaptor protein Shc (Batzer *et al*., 1994; Okutani *et al*., 1994). The SH3 domains of Grb2 are constitutively associated with *S*on *o*f *S*evenless (SOS), an exchange factor of Ras GTPase (Li *et al*., 1993; Rozakis-Adcock *et al*., 1993). Binding of the Grb2–SOS complex to the EGFR places SOS in proximity to Ras, thus leading to GTP-loading of Ras and subsequent activation of Ras effectors, such as Raf kinases and phosphatidylinositol 3-kinase.

Grb2 also is involved in EGF-induced signaling to the actin cytoskeleton (She *et al*., 1997) and EGFR endocytosis (Wang and Moran, 1996; Yamazaki *et al*., 2002; Jiang *et al*., 2003). Although the pathway of Grb2 signaling through Ras is well understood, the molecular mechanisms by which Grb2 controls clathrin-mediated endocytosis of the EGFR remain to be elucidated. The knockdown of Grb2 by small interfering RNA (siRNA) results in dramatic inhibition of the initial steps of EGFR endocytosis (Jiang *et al*., 2003; Huang *et al*., 2004). Grb2 was found to move to coated pits together with activated EGFR (Jiang *et al*., 2003; Stang *et al*., 2004). However, how Grb2 links EGFR to coated pits is unknown. Most likely, Grb2 function is mediated by a protein that binds to the SH3 domains of Grb2. Besides interaction with SOS, Grb2 SH3 domains are capable of association with several proteins, including dynamin and Cbl (c-Cbl, Cbl-b, and Cbl-3) (Meisner *et al*., 1995; Kranenburg *et al*., 1999; Courbard *et al*., 2002), both implicated in the regulation of EGFR endocytosis.

Dynamin is a component of the general clathrin-coated pit machinery and is necessary for endocytosis of all types of cargo internalized through coated pits. Recently, however, it has been proposed that dynamin can be phosphorylated in response to EGF (Ahn *et al*., 2002). The exact role of this phosphorylation and how this dynamin modification contributes specifically to EGFR endocytosis are unclear. Cbl is an E3 ubiquitin ligase that has been implicated in EGFR endocytosis and postendocytic trafficking based on the ef-

This article was published online ahead of print in *MBC in Press* (http://www.molbiolcell.org/cgi/doi/10.1091/mbc.E04–09–0832) on January 5, 2005.

[V] The online version of this article contains supplemental material at *MBC Online* (http://www.molbiolcell.org).

Address correspondence to: Alexander Sorkin (alexander.sorkin@uchsc.edu).

fects of Cbl mutant overexpression (Levkowitz *et al*., 1999; Thien and Langdon, 2001; Dikic and Giordano, 2003; Marmor and Yarden, 2004). However, whereas the role of Cbl in intracellular sorting of EGFR is well established, Cbl participation in the first step of endocytosis, internalization through coated pits, is a subject of debate. The demonstration of Cbl localization in coated pits (de Melker *et al*., 2001; Jiang and Sorkin, 2003; Stang *et al*., 2004) and inhibition of EGFR internalization by c-Cbl mutant overexpression suggested that Cbl is involved in this process (Thien *et al*., 2001; Jiang and Sorkin, 2003). However, the observation of normal EGFR internalization in cells derived from c-Cbl knockout mouse argues against the role of Cbl proteins in this step of EGFR trafficking (Duan *et al*., 2003).

c-Cbl and Cbl-b may function in EGFR endocytosis either by 1) ubiquitylation of the EGFR (Levkowitz *et al*., 1999; Thien *et al*., 2001; Jiang and Sorkin, 2003) or 2) linking EGFR to the CIN85/endophilin complex (Soubeyran *et al*., 2002). The relative contribution of these two pathways in EGFR endocytosis is unknown. The analysis of Cbl role in endocytosis is complicated due to the presence of three Cbl proteins that may have redundant functions.

Previous studies analyzed the spatial and temporal control of Grb2 function in signaling and endocytosis by using overexpression of tagged Grb2 and its interactors. Hence, we developed a new approach to analyze Grb2 function. In this approach, endogenous Grb2 is knocked down by siRNA and replaced by either wild-type Grb2 tagged with yellow fluorescent protein (YFP) or a chimeric protein consisting of the SH2 domain of Grb2 linked to a Grb2-interacting protein, thus bypassing the SH3 domain interactions of Grb2. Our analysis using the "knockdown and rescue" approach was focused on Cbl proteins and revealed the key role of Grb2-mediated recruitment of Cbl to the EGFR in the physiological pathway of receptor internalization via coated pits.

MATERIALS AND METHODS

Reagents

Pfu polymerase and the QuikChange site-directed mutagenesis kit were from Stratagene (La Jolla, CA). EGF conjugated with rhodamine (EGF-Rh) was purchased from Molecular Probes (Eugene, OR). Monoclonal antibodies specific to early endosomal antigen (EEA.1), c-Cbl, and PP1 were obtained from BD Transduction Laboratories (San Diego, CA). Monoclonal antibodies to green fluorescent protein (GFP) and EGFR (Ab528) were from Zymed Laboratories (South San Francisco, CA) and American Type Culture Collection (Manassas, VA), respectively. Polyclonal antibody to Grb2 and monoclonal antibody (mAb) P4D1 to ubiquitin were from Santa Cruz Biotechnology (Santa Cruz, CA). Polyclonal antibodies to total and phosphorylated mitogen-activated protein kinase kinase (MEK)1/2 were from Cell Signaling Technology (Beverly, MA). Rabbit serum Ab2913 specific to the intracellular domain of EGFR was kindly provided by Dr. L. Beguinot (DIBIT Rafaele, Milan, Italy).

Plasmid Constructs and Point Mutations

To generate a Grb2 RNA interference (RNAi) construct (U6/Grb2), a DNA oligonucleotide was inserted into the pSilencer1.0-U6 vector (Ambion, Austin, TX) between *Apa*I (blunted) and *Eco*RI sites. The oligonucleotide, which contained the sense strand of human Grb2 DNA fragment 609–627 followed by a 6-nt loop and then antisense strand of the Grb2 fragment and five Ts, is 5′-G CATGTTTCCCCGCAATTAT CAAGAC ATAATTGCGGGGAAACATG CC TTTTT-3′ (forward) and 5′-AATT AAAAA GG CATGTTTCCCCGCAATTAT GTCTTG ATAATTGCGGGGAAACATG C-3′ (reverse).

The Grb2-YFP construct was described previously (Sorkin *et al*., 2000). Six silent mutations without changing the amino acid sequence of Grb2 were generated in the construct by using the QuikChange mutagenesis kit to make Grb2-YFP resistant to the Grb2 siRNA (609–627) silencing. The DNA primer for the mutagenesis is 5′-C GGG CAG ACC GGA ATG TTC CCA CGT AAC TAC GTC ACC CCC GTG-3′ (forward) and 5′-CAC GGG GGT GAC GTA GTT ACG TGG GAA CAT TCC GGT CTG CCC G-3′ (reverse) (underlined are mutated).

To generate YFP-Cbl, full-length c-Cbl was cut from Cbl-YFP (Jiang and Sorkin, 2003) and cloned into pEYFP-C3 by using *Kpn*I and *Bam*HI sites. Cbl-GFP construct was kindly provided by Drs. G. Levkowitz and Y. Yarden (Weizmann Institute of Science, Rehovot, Israel). DNA encoding SH2 domain of human Grb2 (amino acids 53–164) with the stop codon was amplified using *Pfu* polymerase and inserted into pEYFP-C3 vector (BD Biosciences Clontech, Palo Alto, CA) between *Sac*II and *Bam*HI restriction sites, which generated YFP-SH2. To generate YFP-Cbl-SH2 chimeric proteins, DNAs amplified by polymerase chain reaction encoding full-length c-Cbl or c-Cbl fragments were then inserted between *Xho*I and *Kpn*I sites of YFP-SH2 construct. The chimeras of the c-Cbl N terminus with SH2 domains of rat phospholipase Cγ1 (PLCγ1) (provided by Dr. G. Carpenter, Vanderbilt University, Nashville, TN) were made by cloning DNAs encoding N-terminal domain (corresponding to residues 548–661) or both SH2 domains (residues 548–759) of PLCγ1 with the stop codon into pEYFP-C3 between *Sac*II and *Bam*HI restriction sites followed by inserting c-Cbl fragment between *Xho*I and *Kpn*I sites of YFP-PLC/SH2 constructs. All point mutations in the constructs were generated using the QuikChange mutagenesis kit according to the manufacturer's protocol. All constructs and point mutations were verified by automatic dideoxynucleotide sequencing.

Cell Culture and Transfections

HeLa cells were grown in DMEM containing 10% fetal bovine serum, antibiotics, and glutamine. Grb2 siRNA3 duplex (Jiang *et al*., 2003) was resuspended in 1× siRNA universal buffer provided by Dharmacon (Lafayette, CO), to 20 μM before transfection. To knock down Grb2, HeLa cells in 12-well plates were transfected twice with 4 μl of siRNA3 duplex in 3 μl of LipofectAMINE 2000 reagent (Invitrogen, Carlsbad, CA) at 24-h intervals. For mock-transfections control Cy5-siRNA duplex verified for the absence of off-site effects (Dharmacon) was used. siRNA to μ2 subunit of adaptor protein-2 (AP-2) was prepared and transfected precisely as described previously (Motley *et al*., 2003). For the purpose of expressing Cbl-SH2 chimeric proteins in the Grb2-depleted cells, Cbl constructs were cotransfected with siRNA3 in the second siRNA transfection. Cells were plated to new 12-well plates or glass coverslips 24 h before experiments, which were performed 3 or 4 d after the first transfection.

The HeLa/Grb2-YFP cell lines were obtained by cotransfection of cells with U6/Grb2 and silent Grb2-YFP plasmids followed by single cell clone selection with 0.4 mg/ml G418 (Invitrogen).

Immunoprecipitation and Western Blotting

To assay ubiquitylation of EGFR, HeLa cells transiently transfected with Grb2 siRNA and YFP-Cbl-SH2 in 60-mm dishes were pretreated with 20 ng/ml EGF for 2 min at 37°C and washed with Ca^{2+}, Mg^{2+}-free phosphate-buffered saline (CMF-phosphate-buffered saline). Cells were lysed by scraping with a rubber policeman in Triton X-100/glycerol solubilization buffer (TGH) as described previously (Huang *et al*., 2003, 2004) supplemented with 10 mM *N*-ethyl-maleimide and by incubating further for 10 min at 4°C. The lysates were then cleared by centrifugation for 10 min at 14,000 × *g*. EGFR was immunoprecipitated with antibody Ab528. The precipitates were washed twice with TGH buffer supplemented with 100 mM NaCl, and once without NaCl, and then denatured by heating in sample buffer. Immunoprecipitates and supernatants after immunoprecipitation were resolved on 7.5% SDS-PAGE followed by transfer to nitrocellulose membrane and Western blotting with various antibodies followed by species-specific secondary antibodies or protein A (Zymed Laboratories) conjugated with horseradish peroxidase. The enhanced chemiluminescence kit was from Pierce Chemical (Rockford, IL).

To probe for active MEK1/2, cells in 12-well plates were starved overnight (12–16 h) in DMEM supplemented with 1% conditioned medium, treated with 1 ng/ml EGF at 37°C for 0–30 min, lysed with TGH buffer, and resolved on 10% SDS-PAGE followed by transfer to the nitrocellulose membrane and Western blotting.

^{125}I-EGF Internalization

Mouse receptor-grade EGF (Collaborative Research, Bedford, MA) was iodinated using a modified chloramine T method as described previously (Jiang *et al*., 2003). ^{125}I-EGF internalization was measured using 1 ng/ml ^{125}I-EGF and the specific rate constant for internalization, k_e, was calculated as described previously (Jiang *et al*., 2003). A low concentration of ^{125}I-EGF was used to avoid saturation of the internalization machinery.

Fluorescence Microscopy

HeLa cells on glass coverslips were treated with 2 ng/ml EGF-Rh at 37°C for 5 min or with 10 ng/ml EGF-Rh at 4°C for 1 h, washed with ice-cold phosphate buffer saline, and fixed with freshly prepared 4% paraformaldehyde (Electron Microscopy Sciences, Ft. Washington, PA) for 30 min on ice. In some experiments, cells incubated with EGF-Rh were treated with cold 0.2 M acetic acid, pH 4.5, containing 0.5 M NaCl for 2 min before fixation to remove surface-bound EGF-Rh. In other experiments, to estimate colocalization of EGF-Rh with EEA.1, fixed cells were mildly permeabilized using a 30-min incubation in CMF-phosphate-buffered saline containing 0.02% saponin and 0.1% bovine serum albumin (saponin solution) at room temperature, incubated in saponin solution at room temperature for 1 h with mAb to EEA.1, and then incubated for 30 min with the secondary donkey antimouse IgG

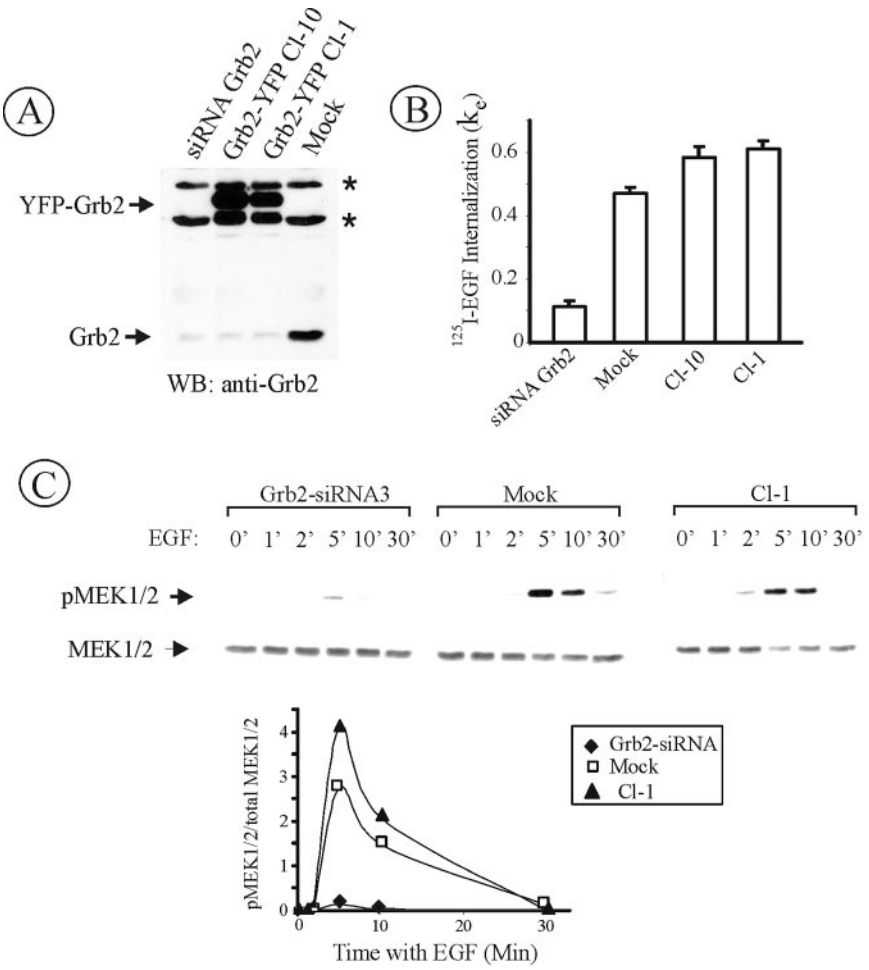


Figure 1. Expression of Grb2-YFP restores EGFR internalization and mitogen-activated protein kinase signaling in Grb2-depleted HeLa cells. (A) Western blot detection of Grb2 in HeLa cells transiently transfected with Grb2 siRNA (siRNA Grb2), transiently mock-transfected (Mock), and HeLa cell lines Cl-10 and Cl-1 stably expressing Grb2-YFP and U6/Grb2-siRNA plasmids. Asterisks show nonspecific bands. (B) ^{125}I-EGF (1 ng/ml) internalization rate constant (k_e) measured in cells described in A. The data represent mean values from three experiments (±SEM). (C) Cells described in A were serum starved overnight and treated with 1 ng/ml EGF at 37°C for the indicated times. Cell lysates were resolved by electrophoresis. Activated MEK1/2 was detected by blotting with antibodies to phosphorylated MEK1/2. The same blots were then reprobed with the antibody to total MEK1/2. The extent of MEK1/2 phosphorylation is expressed as a ratio of phosphorylated to total cellular MEK1/2.

labeled with Cy5 (Jackson ImmunoResearch Laboratories, West Grove, PA). Both primary and secondary antibody solutions were precleared by centrifugation at 100,000 × *g* for 10 min. The coverslips were mounted in Fluoromount-G (Southern Biotechnology, Birmingham, AL) containing 1 mg/ml *para*-phenylenediamine.

Images were acquired using a Marianas imaging workstation (Intelligent Imaging Innovation, Denver, CO). Twelve serial two-dimensional images were recorded at 300-nm intervals. A Z-stack of images obtained was deconvoluted using a nearest neighbor method. The quantitations of EGF-Rh were performed only in experiments where cells were not permeabilized with the detergent because the latter procedure results in a loss of cell-associated EGF-Rh.

The amount of internalized EGF-Rh was quantitated as follows. The background was subtracted from deconvoluted images. The integrated intensity of rhodamine fluorescence in all voxels corresponding to the z-stack of two-dimensional images was calculated for each individual cells. These measurements produced consistently similar values of rhodamine integrated intensity for Grb2-depleted or intact cells in all independent experiments, thus validating the method. The single-cell assay differs from the conventional internalization assay by using ^{125}I-EGF in that the microscopy assay emphasizes the punctate EGF-Rh fluorescence (coated pits, vesicles, and other receptor clusters), whereas it underestimates the diffuse fluorescence of the plasma membrane-bound EGF-Rh due to insufficient sensitivity of imaging systems within the linear range of cooled charged-coupled detectors.

Statistical significance (p value) was assessed using the two-tailed Student's *t* test for unpaired samples.

Time-Lapse Imaging of Grb2-YFP

HeLa/Grb2-YFP cells (Cl-1) were grown on coverslips and assembled into the closed perfusion chamber (Harvard Apparatus, Holliston, MA). The chamber was mounted onto the microscope stage. The time-lapse image acquisition was performed at 33°C through the narrow GFP (excitation 492/10 nm) channel. Typically, 30 images (150-ms integration time) were acquired with 15-s intervals. Binning 2 × 2 mode was used. QuickTime movies were created from the original SlideBook time-lapse images.

RESULTS

Stably Expressed Grb2-YFP Restores EGFR Endocytosis and Signaling in Grb2-depleted Cells

Our previous studies in HeLa and porcine aortic endothelial (PAE) cells demonstrated that knockdown of Grb2 by siRNA results in dramatic inhibition of EGFR internalization (Jiang *et al*., 2003; Huang *et al*., 2004). Hence, we analyzed the mechanisms by which Grb2 regulates EGFR endocytosis. First, we tested whether expression of a Grb2 fusion protein, such as Grb2-YFP, can reverse the effects of Grb2 depletion. To this end, we have chosen a strategy to stably express Grb2-YFP in cells depleted of endogenous Grb2 by siRNA. In this approach, cells constitutively expressing Grb2-YFP at physiological levels can be isolated.

To generate HeLa cells stably expressing Grb2-YFP, endogenous Grb2 was knocked down using vector-based short hairpin RNA (shRNA) with simultaneous expression of Grb2-YFP that has silencing mutations rendering this construct insensitive to shRNA. From the selected clones, two single cell clones (Cl-1 and Cl-10) of HeLa/Grb2-YFP cells with minimal expression of endogenous Grb2 and moderate expression of Grb2-YFP were chosen (Figure 1A). Cl-1 expressed very low amount of Grb2 (<5% of Grb2 level in mock-transfected cells) and a nearly physiological level of Grb2-YFP (~40% higher than the Grb2 level in the parent HeLa cells). This clone was used in most of the subsequent experiments.

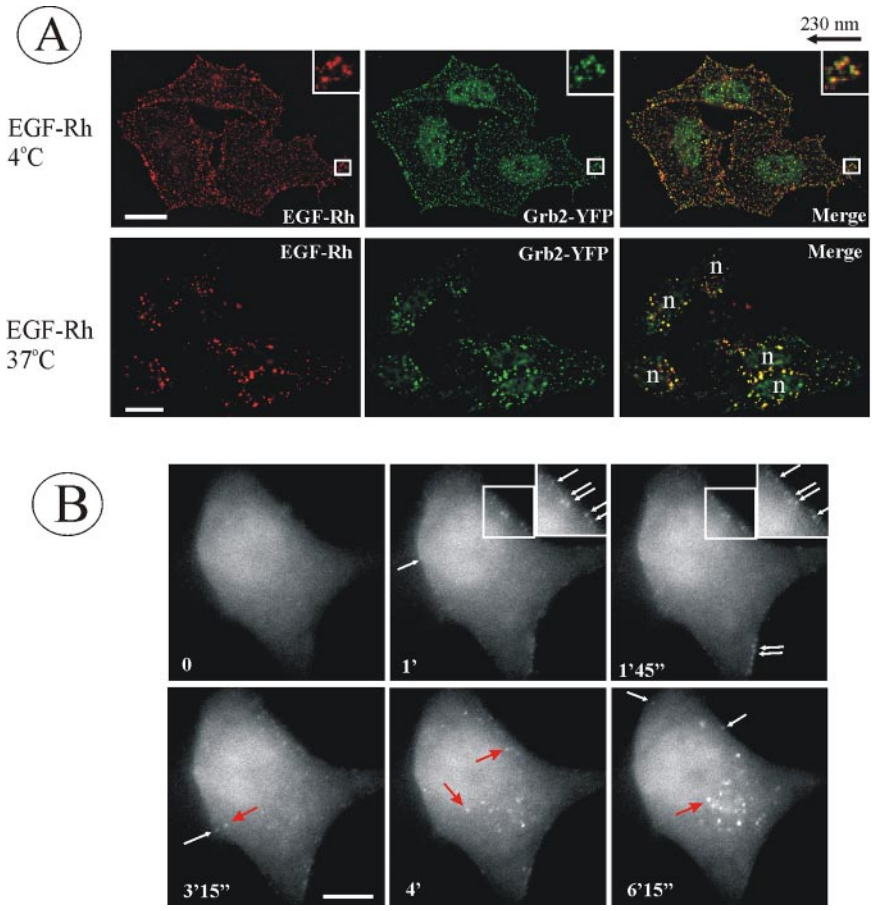


Figure 2. Visualization of Grb2-YFP in HeLa/Grb2-YFP *Cl-1* cells. (A) Cl-1 cells were treated with 10 ng/ml EGF-Rh at 4°C for 1 h or with 2 ng/ml EGF-Rh at 37°C for 5 min and fixed. A z-stack of images was acquired through the Cy3 (red) and YFP (green) channels and deconvoluted. YFP and Cy3 images representing individual optical sections were merged after adjustment of both fluorescence signals to similar levels (Merge). Insets represent high-magnification images of the regions shown by white rectangles. In the merged inset image, YFP image was shifted approximately by 230 nm to the left relative to rhodamine images to clearly assess the colocalization of Grb2-YFP with EGF-Rh. n, cell nuclei. Bars, 10 μm. (B) Cells were incubated with 100 ng/ml EGF at 33°C, and images were acquired at 15-s intervals. This high concentration of EGF was used because of the significant dilution of EGF during perfusion of the microscope chamber. Times after EGF addition are indicated in the left bottom corner of the images. White arrows show Grb2-YFP localized to endocytic coats and forming vesicles at cell edges, whereas red arrows show Grb2-YFP localized in endosomes at some distance from cell edges or in the perinuclear area of the cell. The corresponding Quicktime movie is presented in Supplemental Materials. Bar, 10 μm.

Both Cl-1 and Cl-10 displayed high rates of ^{125}I-EGF internalization, indicating that Grb2-YFP fully rescued the inhibitory effects of Grb2 knockdown on EGFR endocytosis (Figure 1B). Depletion of Grb2 by siRNA also caused dramatic inhibition of EGF-induced MEK1/2 activation (Figure 1C). MEK1/2 activity was, however, fully restored in Cl-1 cells (Figure 1C). Interestingly, ERK1/2 activity was inhibited only by 50% in cells depleted of Grb2 and also fully rescued by Grb2-YFP (our unpublished data). The significant activation of ERK1/2 in cells with very low MEK1/2 activity can be due to substantial molar excess of MEK1/2 over ERK1/2 in the cells and a 1000-fold increase of ERK kinase activity by MEK1/2, leading to dramatic signal amplification at this step of the ERK activation cascade (reviewed in Pearson *et al*., 2001). In fact, binding of EGF to a very small pool of EGFR was sufficient for the full activation of ERK1/2 in several cell types (Soler *et al*., 1994a; Johannessen *et al*., 2000).

The effective endocytosis of EGFR also could be observed in HeLa/Grb2-YFP cells by fluorescence microscopy by using an EGF-Rh conjugate. When Cl-1 cells were incubated with EGF-Rh at 4°C, conditions that allow recruitment of the EGFR into coated pits but restrict endocytosis, Grb2-YFP dots were highly colocalized with EGF-Rh dots at the plasma membrane (Figure 2A). Previous studies demonstrated that these dots represent clathrin-coated pits (Jiang *et al*., 2003). When cells were incubated with EGF-Rh at 37°C, EGF-Rh and Grb2-YFP were accumulated in the peripheral and perinuclear endosomes (Figure 2A).

To examine localization of Grb2-YFP in living cells, time-lapse imaging of EGF-stimulated cells was performed (Figure 2B) (see Quicktime movie in Supplemental Materials). In unstimulated cells, Grb2-YFP was diffusely distributed in the cytosol and nucleus. Similar pattern of localization was observed for endogenous Grb2 by using immunofluorescence staining (Sorkin *et al*., 2000). Rapidly after EGF stimulation, Grb2-YFP began to concentrate in small dots at the edges of the cell. Further incubation at 37°C led to accumulation of Grb2-YFP in endosomes located mostly in the perinuclear area of the cell. Together, microscopy analyses of fixed and living cells demonstrated that Grb2-YFP accompanies EGFR at the plasma membrane and during the passage of receptors through endocytic compartments. These data also illustrate that the dynamic localization of Grb2-YFP can be analyzed in cells, in which this fusion protein is expressed at physiological levels.

Cbl Overexpression Does Not Rescue EGF Endocytosis in Grb2-depleted Cells

Stemming from our findings that Grb2 is essential for EGFR internalization in HeLa cells and that Grb2-YFP can functionally replace endogenous Grb2, we further investigated the mechanisms by which Grb2 controls EGFR internalization. Because Cbl proteins have been implicated in EGFR internalization (Levkowitz *et al*., 1999; Thien *et al*., 2001; Waterman *et al*., 2002; Jiang and Sorkin, 2003), we tested whether overexpression of c-Cbl can overcome the inhibition of endocytosis imposed by Grb2 knockdown. In these experiments, cells were cotransfected with Grb2-targeted siRNA and a YFP-Cbl construct (Figure 3A). The endocytosis of EGFR was analyzed using a fluorescence microscopy assay by comparing EGF-Rh endocytosis in Grb2-depleted

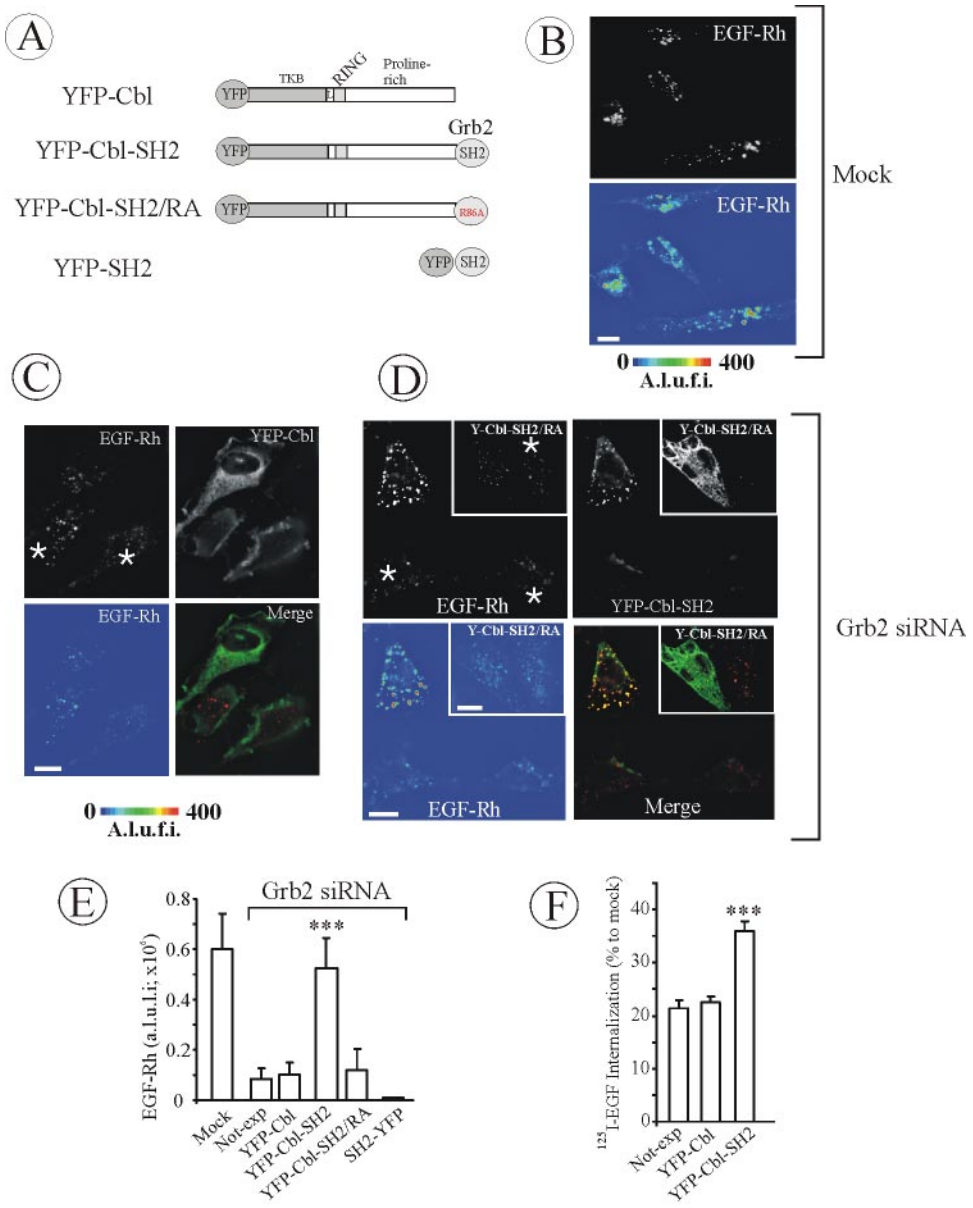


Figure 3. YFP-Cbl-SH2 chimera rescues EGFR internalization in HeLa cells depleted of Grb2. (A) Schematic representation of Cbl-Grb2 chimeric proteins fused to YFP. The YFP-Cbl-SH2 chimera consists of YFP, residues 1–906 of c-Cbl, and the SH2 domain of Grb2. YFP-Cbl-SH2/RA is a mutant YFP-Cbl-SH2 chimera with the substitution of Arg86 by alanine in the Grb2 SH2 domain. YFP-SH2 is the SH2 domain of Grb2 fused to YFP. (B–D) HeLa cells were mock transfected (B) or transiently transfected (C and D) with Grb2 siRNA. The cells also were transfected with YFP-Cbl (C), YFP-Cbl-SH2 (D), or YFP-Cbl-SH2/RA (D, insets). The cells were incubated with 2 ng/ml EGF-Rh for 5 min at 37°C and fixed. A z-stack of optical sections was acquired through the Cy3 (red) and YFP (green) channels and deconvoluted. YFP and Cy3 images representing individual optical sections were merged after adjustment of the YFP signal (Merge). EGF-Rh images also are displayed in quantitative pseudocolor using the same scale for all images. A.l.u.f.i., arbitrary linear units of fluorescence intensity. White asterisks show positions of Grb2-depleted cells that do not express YFP fusion proteins (nonexp). Some of these cells can be seen by their autofluorescence in the YFP filter channel. Bars, 10 μm. (E) Quantification of the amount of internalized EGF-Rh from experiments performed as described in B–D. In experiments with all YFP-tagged proteins, the values of integrated intensity did not depend on the expression levels of these fusion proteins. The error bars represent SEs (n = 20). ***$p < 0.0001$ compared with cells depleted of Grb2 and not expressing YFP-tagged constructs (nonexp). (F) HeLa cells were transfected with control siRNA (mock) or Grb2 siRNA. The cells were also transfected with YFP-Cbl or YFP-Cbl-SH2 as described in D. Internalization rates of ^{125}I-EGF were measured as in Figure 1B.

cells expressing or not expressing YFP-Cbl. The cells were incubated with 2 ng/ml EGF-Rh for 5 min at 37°C, acid washed (pH 4.5) to remove surface EGF-Rh, and analyzed by three-dimensional (3-D) deconvolution microscopy as described in *Materials and Methods*. These conditions are optimized for monitoring the clathrin-mediated pathway of EGFR internalization. For instance, endocytosis of EGF-Rh measured using an identical assay was severely inhibited by siRNA knockdown of clathrin heavy chain (Huang *et al.*, 2004).

As shown in Figure 3B, EGF-Rh was rapidly internalized and accumulated in early/intermediate endosomes in the

perinuclear area of control HeLa cells (processed with all siRNA transfection reagents). In Grb2-depleted cells, a small amount of EGF-Rh was present in the perinuclear vesicular structures (presumably, endosomes) (see cells not expressing YFP marked by asterisks in Figure 3, C and D), and it was virtually impossible to find cells with the significant amount of internalized EGF-Rh. The residual internalization in Grb2-depleted cells was probably due to either Grb2-independent pathways of internalization or a small pool of Grb2 remaining in the cells. The pattern of localization and the intensity of EGF-Rh fluorescence was essentially the same in cells expressing or not expressing YFP-Cbl (Figure 3C) or Cbl-GFP fusion proteins (our unpublished data). These data suggest that c-Cbl overexpression did not functionally compensate for the absence of Grb2.

Recruitment of Cbl to the EGFR through the SH2 Domain of Grb2 Restores EGFR Endocytosis in Grb2-depleted Cells

Our previous study suggested the importance of Cbl binding to the EGFR through the interaction of its carboxy terminus with the SH3-containing proteins, such as Grb2 (Jiang and Sorkin, 2003). Thus, we reasoned that to functionally rescue the effects of Grb2 knockdown, Cbl must be situated on the receptor molecule in a way that would mimic Cbl position when it is bound to Grb2. Therefore, a chimeric protein containing full-length c-Cbl fused at the carboxy terminus to the SH2 domain of Grb2 (YFP-Cbl-SH2) was constructed (Figure 3A). Such a chimera was expected to bind to Grb2 binding sites of the EGFR.

Surprisingly, when expressed in Grb2-depleted cells, YFP-Cbl-SH2 caused significantly larger accumulation of EGF-Rh in endosomes compared with Grb2-depleted cells that did not express the chimeric protein (Figure 3D). YFP-Cbl-SH2 was highly colocalized with EGF-Rh, suggesting that the chimera efficiently binds to the EGFR. Mutational inactivation of the Grb2 SH2 domain in the YFP-Cbl-SH2 construct (R86A mutation) resulted in a chimeric protein that failed to restore EGF-Rh endocytosis in Grb2-depleted cells (Figure 3D, insets). This observation established that the rescue effect of YFP-Cbl-SH2 chimera requires the functional SH2 domain of Grb2.

Quantitative analysis revealed that the amount of internalized EGF-Rh in Grb2-depleted, YFP-Cbl-SH2–expressing cells was significantly higher than the amount of EGF-Rh in neighboring cells that did not express the YFP-Cbl-SH2 chimera or in independent cultures of Grb2-depleted, DNA mock-transfected cells (Figure 3E). Although the extent of Grb2 depletion in each individual cell could not be controlled by immunostaining because of the loss of EGF-Rh due to cell permeabilization, very small variation in the amount of EGF-Rh per cell within the population of randomly chosen Grb2-depleted cells (Figure 3E) suggested that EGF-Rh endocytosis was substantially inhibited in all cells in the population.

The extent of EGF-Rh endocytosis in Grb2-depleted and YFP-Cbl-SH2–expressing cells was similar to that in cells that were not depleted of Grb2 (Figure 3E). Interestingly, control experiments, in which Grb2 SH2 domain tagged with YFP (Figure 3A) (Jiang *et al*., 2003) was expressed in Grb2-depleted cells, revealed that this fusion protein not only failed to rescue EGFR endocytosis but caused complete disappearance of detectable internalized EGF-Rh (Figure 3E). This observation implies that the residual endocytosis in Grb2-depleted cells is likely to be dependent on Grb2 binding sites in the EGFR, which are blocked by overexpression of SH2-YFP.

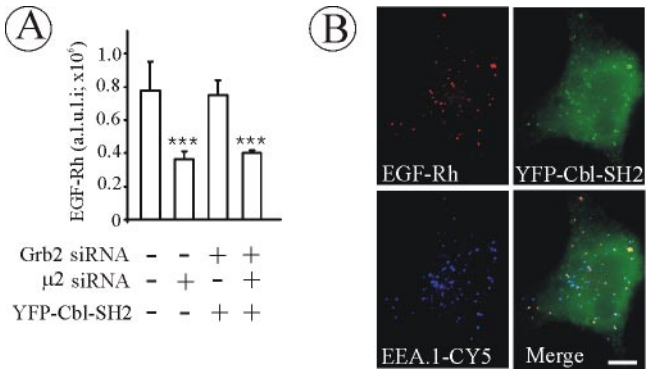


Figure 4. EGF-Rh is internalized in AP-2–dependent manner and accumulates in early endosomes in cells depleted of Grb2 and expressing YFP-Cbl-SH2 chimera. (A) HeLa cells were mock transfected or transfected with Grb2 siRNA or μ2 siRNA. The cells were also transfected with YFP-Cbl-SH2 as described in Figure 3D. The incubations with EGF-Rh and quantitations of the amount of EGF-Rh in endosomes were performed as described in Figure 3, B–E. The error bars represent SEs ($n = 10$). ***$p < 0.0001$ compared with cells depleted of μ2. (B) HeLa cells were transfected with Grb2 siRNA and YFP-Cbl-SH2, incubated with EGF-Rh as described in D, fixed, permeabilized, and stained with a mAb to EEA.1 followed by secondary donkey anti-mouse IgG conjugated with Cy5. A z-stack of optical sections was acquired through the Cy3, Cy5, and YFP channels and deconvoluted. Cy5, YFP, and Cy3 images representing individual optical sections cells were merged (Merge). Yellow signifies colocalization of rhodamine and YFP labels; cyan of rhodamine and Cy5, and white of all three labels. Bar, 10 μm.

In addition to the single-cell internalization assay, ^{125}I-EGF endocytosis was measured in total cell populations of Grb2-depleted cells as described in Figure 1B. As shown in Figure 3F, the apparent rate of internalization was increased by ~65% in cells transfected with YFP-Cbl-SH2, whereas it was not increased by expression of YFP-Cbl compared with Grb2-depleted cells. Such partial increase was expected because the efficiency of DNA plasmid transfection under conditions of siRNA/DNA cotransfection was significantly lower (15–20%) compared with that of siRNA transfection (essentially 100%). This resulted in YFP-Cbl-SH2 expression and therefore rescue of endocytosis in only a small pool of Grb2-depleted cells. Because the data obtained in two internalization assays correlated well and because the conventional biochemical assay is less reliable due to difficulties in matching expression levels and transfection efficiencies of different Cbl-SH2 fusion constructs, subsequent experiments relied on the quantitative single-cell assay.

The data in Figure 3 demonstrated that the YFP-Cbl-SH2 chimera that is capable of binding to Grb2 interaction sites of the EGFR restores EGFR endocytosis in Grb2-depleted cells. To test whether this restored endocytosis is mediated by clathrin-associated machinery, the effect of depletion of clathrin adaptor complex AP-2 on EGFR endocytosis in Grb2-depleted/YFP-Grb2-SH2 chimera-rescued cells was tested using siRNA targeting μ2 subunit of AP-2 (Motley *et al*., 2003). As shown in Figure 4A, μ2 siRNA inhibited EGF-Rh endocytosis to the same extent in mock-transfected (control siRNA and DNA) and Grb2-depleted/YFP-Cbl-SH2–rescued cells. Furthermore, EGF-Rh was well colocalized with EEA.1 in YFP-Cbl-SH2–expressing cells (Figure 4B), confirming that the vesicular-shape compartments containing EGF-Rh represent early/sorting endosomes. These data imply that the pathways of EGFR endocytosis in YFP-

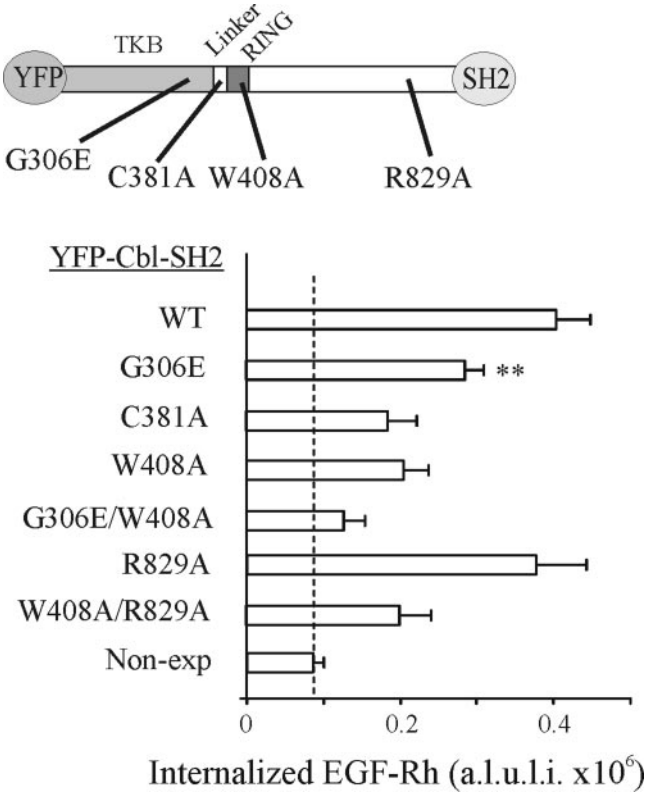


Figure 5. Point mutations in the YFP-Cbl-SH2 chimera affect endocytosis rescue capacity of the chimera in cells depleted of Grb2. HeLa cells were transiently transfected with Grb2 siRNA and either with YFP-Cbl-SH2 (WT) or its mutants schematically shown on the top. The cells were incubated with EGF-Rh, fixed, and imaged as described in Figure 3. The quantification of the amount of internalized EGF-Rh from experiments was performed as in Figure 3E. Cells transfected with Grb2 siRNA and not expressing YFP fusion proteins are designated as nonexp. In experiments with all YFP-tagged proteins, the values of integrated intensity did not depend on the expression levels of these fusion proteins. The error bars represent SEs ($n = 20$). **$p < 0.002$ compared with cells depleted of Grb2 and expressing YFP-Cbl-SH2 (*WT*). A.l.u.f.i., arbitrary linear units of fluorescence intensity.

Cbl-SH2–rescued cells are similar to those pathways normally utilized by EGFR in intact cells.

Cbl RING Finger Domain Is Critical for the Rescue Effects of the Cbl-Grb2 Chimera

Figure 3 shows that the recruitment of c-Cbl to the EGFR via receptor Grb2 binding sites is sufficient to compensate for the inhibitory effect of Grb2 knockdown on receptor endocytosis. To examine which functions of YFP-Cbl-SH2 chimera are responsible for the rescue effects, several point mutations in the SH2-like/phosphotyrosine binding (PTB) domain (G306E), linker (C381), RING domain (W408A) (Levkowitz *et al*., 1999; Thien *et al*., 2001), and CIN85 binding site (R829A) (Kowanetz *et al*., 2003) were made in the chimera, and the resulting mutants were tested for the ability to rescue EGF-Rh endocytosis in Grb2-depleted cells as described in Figure 3. The quantitative analysis of 3-D images revealed that both mutations that abolish the activity of RING domain (C381A and W408A) resulted in a significantly reduced ability to support EGF-Rh endocytosis compared with "wild-type" Cbl-SH2 chimera (Figure 5). Very minimal, if any, colocalization of EGF-Rh with EEA.1 was detected in cells expressing YFP-Cbl/C381A-SH2 or YFP-Cbl/W408A-SH2 fusion proteins (our unpublished data). In contrast, mutation of CIN85 binding site did not affect the internalization rescue capacity of YFP-Cbl-SH2. Interestingly, a mutant with defective SH2-like domain (G306E) displayed intermediate rescue capacity (Figure 5). These data indicate that the RING domain is critical for the Cbl function in endocytosis. However, inactivation of the RING domain did not abolish EGFR endocytosis to the level observed in Grb2-depleted cells, suggesting that interactions of other parts of Cbl molecule may support endocytosis in the absence of RING function.

To test whether expression of YFP-Cbl-SH2 chimera results in ubiquitylation of EGFR, EGFR was immunoprecipitated from Grb2-depleted cells that were DNA mock transfected or transfected with YFP-Cbl, YFP-Cbl-SH2, or YFP-Cbl/W408A-SH2. EGFR ubiquitylation was not detected in unstimulated cells regardless of whether cells were transfected or not with YFP constructs. EGF stimulation of mock-transfected cells resulted in the appearance of a characteristic smear of high-molecular-weight ubiquitin immunoreactivity in EGFR immunoprecipitates (Figure 6A). The ubiquitylation of EGFR was, however, dramatically inhibited in cells depleted of Grb2. These data demonstrate that Grb2 is essential for Cbl-mediated ubiquitylation of EGFR.

Expression of YFP-Cbl-SH2 chimera in Grb2-depleted cells recovered EGFR ubiquitylation (Figure 6A). Consistently, Cbl-Grb2 chimeric proteins were effectively coprecipitated with EGFR due to the interaction of the Grb2 SH2 domain with the receptor (Figure 6B). The ubiquitylation-rescue effect was absolutely dependent on the presence of the functional RING domain because W408A (Figure 6A) or C381A mutants (our unpublished data) of the chimera did not restore EGFR ubiquitylation in Grb2-depleted cells. In contrast, EGFR ubiquitylation in Grb2-depleted cells was not rescued by overexpression of YFP-Cbl as assessed by quantitations of the amounts of ubiquitin immunoreactivity normalized to the EGFR immunoreactivity in EGFR immunoprecipitates (Figure 6, A and C). This conclusion is supported by the observation of negligible coimmunoprecipitation of YFP-Cbl with EGFR in the absence of Grb2 (Figure 6B).

Cbl Amino-Terminal Domain Is Sufficient for Mediating Grb2-dependent Endocytosis of EGFR

The data of Figures 5 and 6 demonstrated the importance of RING domain function in the endocytosis rescue capacity of Cbl-Grb2 chimeras. To test whether the functional RING domain is sufficient to recover EGFR endocytosis in Grb2-depleted cells, a fusion of the Grb2 SH2 domain and the amino terminus of c-Cbl (residues 1–450) encompassing the tyrosine kinase binding (TKB) linker and RING domains was generated (YFP-C′450-SH2; Figure 7A). Several amino acid residues in the TKB domain are engaged in the intramolecular interactions with the linker and RING domains ensuring proper folding of the RING domain (Zheng *et al*., 2000). Figure 7B shows that EGF-Rh was efficiently internalized in Grb2-depleted cells expressing this chimeric protein. The pattern of EGF-Rh localization was indistinguishable from that observed in cells expressing a chimeric protein containing the full-length c-Cbl. EGF-Rh was well colocalized with EEA.1 in endosomes of cells expressing YFP-C′450-SH2 (Figure 7D). The amounts of endosomal EGF-Rh were similar in cells expressing YFP-C′450-SH2 and YFP-Cbl-SH2 proteins (Figure 7C). Measurements of ^{125}I-EGF uptake performed as described in Figure 3F demonstrated essentially the same rescue effect (~65%) of YFP-C′450-SH2

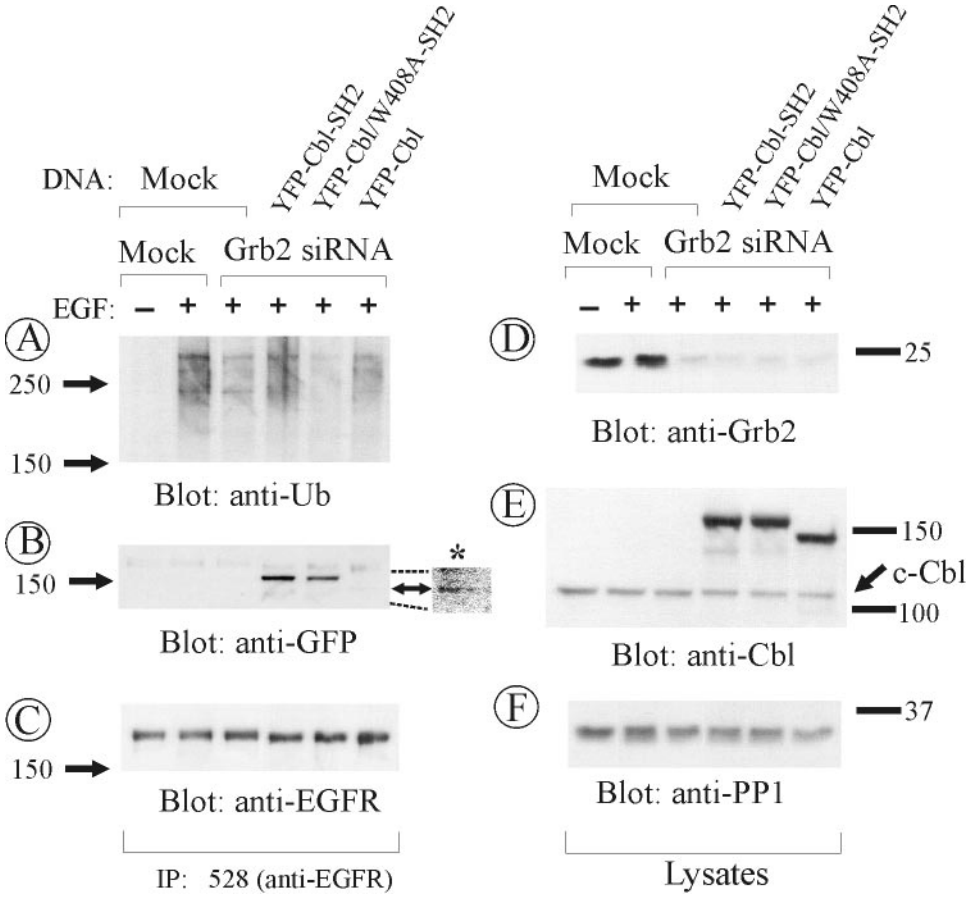


Figure 6. YFP-Cbl-SH2 chimeras bind EGFR and rescues ubiquitination of EGFR in Grb2-depleted cells. HeLa cells were transiently mock siRNA transfected or transfected with Grb2 siRNA. The cells also were mock DNA transfected or transfected with YFP-Cbl-SH2 chimera, mutant W408A of this chimera (see Figure 4), or YFP-Cbl. The cells were incubated with 20 ng/ml EGF for 2 min at 37°C and lysed. EGFR was immunoprecipitated (IP) with antibody 528. Immunoprecipitates and lysates were resolved on 7.5% SDS-PAGE and probed by Western blotting with antibody P4D1 to ubiquitin (A), GFP (B), antibody 2913 to EGFR (C), Grb2 (D), c-Cbl (E), and PP1 (F, nonspecific control). Asterisk in B marks a portion of the blot that was overexposed to demonstrate the presence of a small amount of YFP-Cbl in EGFR immunoprecipitates.

and YFP-Cbl-SH2 proteins (our unpublished data). Together, the data in Figure 7 suggested that the amino terminus of Cbl recruited to the Grb2 binding sites of the EGFR is sufficient to reverse the inhibition of EGFR internalization caused by Grb2 knockdown.

To test whether SH2-like, linker, and RING domains of Cbl are important for the rescue ability of the Cbl amino terminus, corresponding mutants of YFP-C′450-SH2 were prepared (Figure 7A). Expression of YFP-C′450-SH2 with mutations inactivating RING domain failed to restore EGFR endocytosis (Figure 7, B and C). As in experiments with the full-length Cbl-Grb2 chimeras (Figure 5), G306E mutation in the YFP-C′450-SH2 partially abolished the rescue effect of this chimera (Figure 7C).

Thus, the data in Figure 7 demonstrated that the association of the amino-terminal domain of c-Cbl containing functional RING domain with the EGFR is necessary and sufficient to restore EGFR endocytosis in cells depleted of Grb2. The SH2/-like domain of Cbl may be necessary for the full effectiveness of the RING domain.

To test whether recruitment of the Cbl amino-terminal domain via an SH2 domain of EGFR-binding proteins other than Grb2 also can restore endocytosis in Grb2-depleted cells, we took advantage from the observation of the robust binding of the N-terminal SH2 domain of PLCγ1 to the EGFR, which is maintained during EGFR endocytosis (Chattopadhyay *et al*., 1999; Matsuda *et al*., 2001; Wang *et al*., 2001; Wang and Wang, 2003). The SH2 domain of Grb2 in YFP-C′450-SH2 was replaced by either single N-SH2 or a tandem of two SH2 domains of PLCγ1 to generate YFP-C′450-PLC.N/SH2 and C′450-PLC.NC/SH2, respectively (Figure 8A). Expression of these chimeric proteins restored EGF-Rh endocytosis in Grb2-depleted cells (Figure 8, B and C). Both Cbl-PLCγ1 chimeras were colocalized with endosomal EGF-Rh, indicative of high affinity and stable association with activated EGFR (Figure 8B). These data suggest that the strength and longevity of association with EGFR are important for the endocytosis-rescue activity of the Cbl amino terminus. Because SH2 domains of PLCγ1 can bind to a number of phosphotyrosines of EGFR, including Tyr1068 (Soler *et al*., 1994b; Chattopadhyay *et al*., 1999), it is difficult to determine whether a particular architecture of the EGFR:chimera complex is beneficial for the endocytosis-rescue effects.

The Interaction of Cbl with CIN85 Has a Redundant Role in EGFR Endocytosis

To test whether Cbl interaction with CIN85 has a role in Grb2-mediated EGFR endocytosis, we tested whether the carboxy-terminal domain of Cbl, responsible for interaction

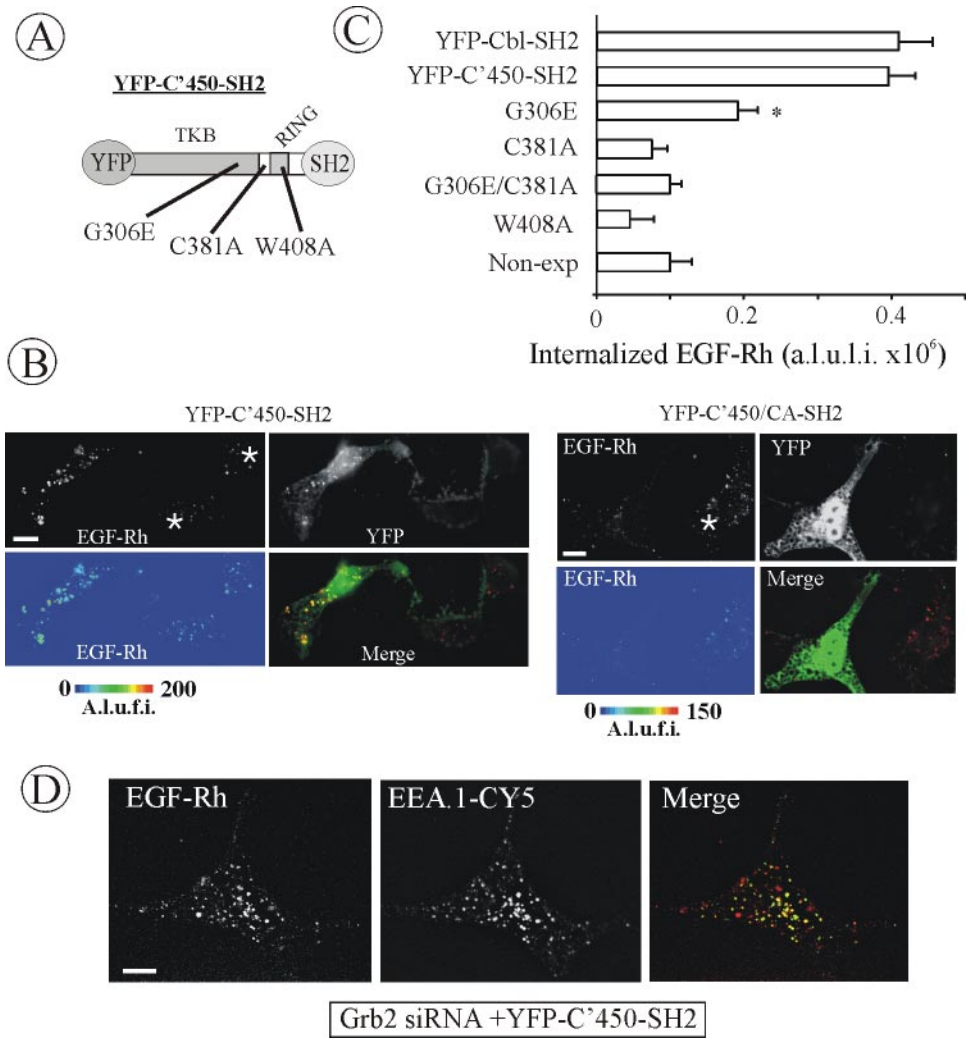


Figure 7. Grb2 chimera with amino terminus of Cbl restores EGFR internalization in Grb2-depleted cells. (A) Schematic representation of YFP-C′450-SH2 chimeric protein. YFP-C′450-SH2 contains c-Cbl fragment encompassing residues 1–450 flanked with YFP and the SH2 domain of Grb2. Mutations made in the construct are indicated. For example, YFP-C′450/CA-SH2 is a mutant of YFP-C′450-SH2 with the substitution of Cys381 by alanine. (B) HeLa cells were transfected with Grb2 siRNA and YFP-C′450-SH2, YFP-C′450/CA-SH2 or other mutants (our unpublished data). The cells were treated with EGF-Rh, fixed, and imaged as described in Figure 3. YFP and Cy3 images representing individual optical sections were merged after adjustment of the YFP signal (Merge). EGF-Rh images also are displayed in quantitative pseudocolor by using the same scale for both images. A.l.u.f.i., arbitrary linear units of fluorescence intensity. White asterisks show positions of Grb2-depleted cells that do not express YFP fusion proteins (nonexp). Bars, 10 μm. (C) Quantification of the amount of internalized EGF-Rh from experiments performed as described in B. In experiments with all YFP-tagged proteins, the values of integrated intensity did not depend on the expression levels of these fusion proteins. The error bars represent SEs ($n = 15–20$). *$p < 0.05$ compared with cells depleted of Grb2 and not expressing YFP-tagged constructs (nonexp). (D) HeLa cells transfected with Grb2 siRNA and YFP-C′450-SH2 were incubated with EGF-Rh, fixed, and processed for immunostaining with EEA.1 antibody as in Figure 3F. The z-stack of optical sections was acquired through the Cy3 (red), Cy5 (green), and YFP (our unpublished data) channels and deconvoluted. Cy5 and Cy3 images representing individual optical sections were merged (Merge). Yellow signifies colocalization of rhodamine and Cy5 labels. Bar, 10 μm.

with CIN85, is capable of restoring EGF-Rh endocytosis in Grb2-depleted cells. To this end, a YFP-tagged fusion of the Cbl fragment encompassing residues 435–906 with the Grb2 SH2 domain was generated (Figure 9A). The expression of this chimera in Grb2-depleted cells resulted in a slightly larger accumulation of EGF-Rh in endosomes as compared with non-YFP–expressing cells, suggesting a partial rescue effect. However, the extent of the rescue of EGFR endocytosis was 4 times smaller than that of the rescue achieved in cells expressing a RING-containing chimeric protein (YFP-C′450-SH2) (Figure 9, B and C). The reduced rate of endocytosis also was suggested by the observation that EGF-Rh was accumulated in EEA.1-containing endosomes in cells expressing the YFP-N′435-SH2 chimera only after 15 min of continuous endocytosis (Figure 9D). In contrast, 5 min of internalization was sufficient for EGF-Rh to reach large perinuclear EEA.1-positive endosomes in cells rescued by expression of RING-containing chimeras (Figure 7C).

The partial rescue of EGFR endocytosis by YFP-N′435-SH2 was dependent on the interaction of this chimeric protein with the SH3 domain of CIN85 because mutation of Arg829, a key residue in the CIN85 interaction motif in c-Cbl (Kowanetz *et al.*, 2003), abolished the rescue effect of the YFP-N′435-SH2 chimera (Figure 9, B and C). However, in

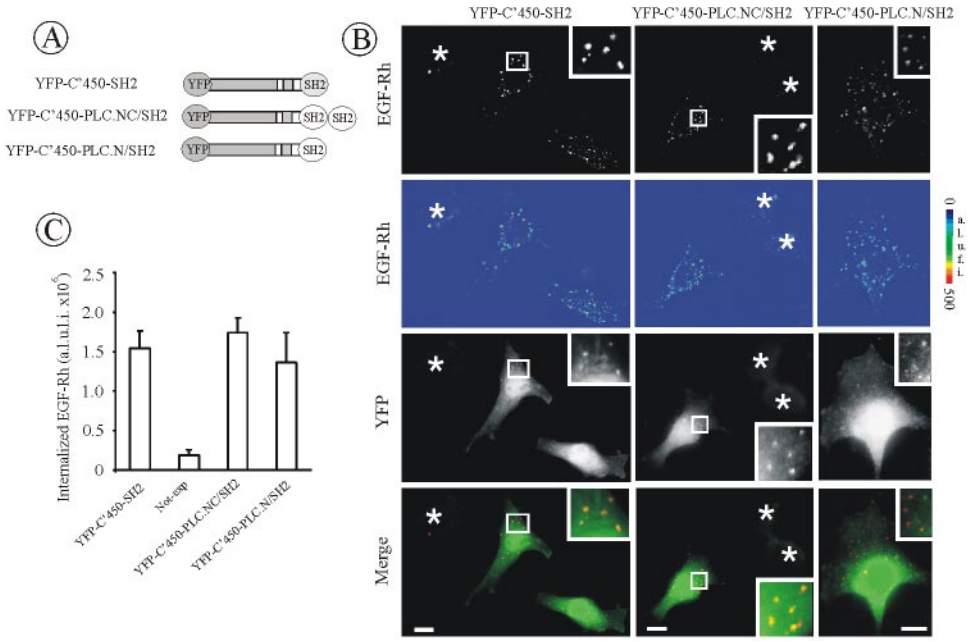


Figure 8. Chimera consisting of PLCγ1 SH2 domains and the amino terminus of Cbl restores EGFR internalization in Grb2-depleted cells. (A) Schematic representation of YFP-C′450-SH2 chimeric proteins. YFP-C′450-PLC.N/SH2 and YFP-C′450-PLC.NC/SH2 contain c-Cbl fragment encompassing residues 1–450 flanked with YFP and, respectively, the single N-SH2 domain or a tandem of two SH2 domains of PLCγ1. (B) HeLa cells were cotransfected with Grb2 siRNA and either YFP-C′450-SH2, YFP-C′450-PLC.N/SH2, or YFP-C′450-PLC.NC/SH2. The cells were treated with EGF-Rh, fixed, and imaged as described in Figure 3. YFP and Cy3 images representing individual optical sections were merged after adjustment of the YFP signal (Merge). EGF-Rh images also are displayed in quantitative pseudocolor by using the same scale for both images. Insets represent high-magnification images of the regions shown by white rectangles and demonstrate localization of the chimeric proteins in endosomes containing EGF-Rh. A.l.u.f.i., arbitrary linear units of fluorescence intensity. Bars, 10 μm. (C) Quantification of the amount of internalized EGF-Rh from experiments similar to those presented in B. The error bars represent SEs (n = 10).

agreement with Figure 4, R829A mutation had no influence on the endocytosis rescue capacity of the full-length Cbl chimera in Grb2-depleted cells (Figures 9C). These data suggest that Arg829-mediated interactions are not essential for Cbl function in EGF-Rh endocytosis but may participate in endocytosis in the absence of RING domain activity of Cbl proteins associated with the EGFR.

DISCUSSION

In this study, we exploited the siRNA knockdown and rescue approach to confirm the role of Grb2 in EGFR internalization and to further investigate the mechanisms by which Grb2 participates in this process. At least in two types of cells, HeLa and PAE cells, Grb2 is essential for clathrin-mediated pathway of EGFR internalization (Jiang *et al*., 2003; Huang *et al*., 2004; this study). In both cell types, this conclusion was reached under conditions of moderate EGFR expression and by using low physiological concentrations of EGF. In NIH 3T3 and B82 mouse fibroblast cell lines, expressing transfected EGFR, Grb2 depletion by siRNA resulted in 65 and 40% reduction of EGF internalization rates, respectively (our unpublished data). Elimination of Grb2 binding sites in the EGFR by point mutations or large deletions did not affect receptor internalization in NIH 3T3 and B82 cells (Chen *et al*., 1989; Chang *et al*., 1993; Sorkina *et al*., 2002). Therefore, there might be other physiological and/or compensatory mechanisms of EGFR endocytosis that do not rely on Grb2. These alternative pathways can be clathrin dependent and independent. For example, it has not been tested whether endocytosis of EGFR mutants lacking Grb2 binding sites in NIH 3T3 and B82 cells is mediated by clathrin-coated pits. For some truncated EGFR mutants, such as C′1022, the kinetics and EGFR kinase dependence of internalization is different from that of wild-type EGFR, which indicates that the mechanism of endocytosis of such truncated mutants may not be physiological (Chen *et al*., 1989; our unpublished data). In addition, in the absence of Grb2 binding sites in EGFR mutants, other phosphotyrosines in the EGFR may bind Grb2 and support endocytosis, especially in cells overexpressing these EGFR mutants. Finally, the relative contribution of Grb2-dependent and other pathways of clathrin-dependent and -independent endocytosis may be cell specific.

The mechanisms of Grb2-dependent EGFR endocytosis have been difficult to analyze because Grb2 is indispensable for cell growth and survival. The importance of Grb2 is illustrated by early embryonic lethality of Grb2 knockout mice that made it impossible to generate mouse embryonic fibroblast lines from Grb2−/− animals (Cheng *et al*., 1998). Likewise, we were unable to develop stable cell lines depleted of Grb2 or expressing Grb2-YFP or its mutants. The cells constitutively expressing Grb2-YFP could be generated only when endogenous Grb2 was eliminated by shRNA, and Grb2-YFP was expressed at physiological levels (Figure 1). The development of cell lines expressing fluorescently tagged Grb2, in which the effects of siRNA Grb2 knockdown on endocytosis and signaling were fully reversed, allowed demonstration the specificity of Grb2 siRNA effects.

The data obtained with HeLa/Grb2-YFP cells have additional implications. Over the years, numerous studies, including from our laboratory, have analyzed localization and

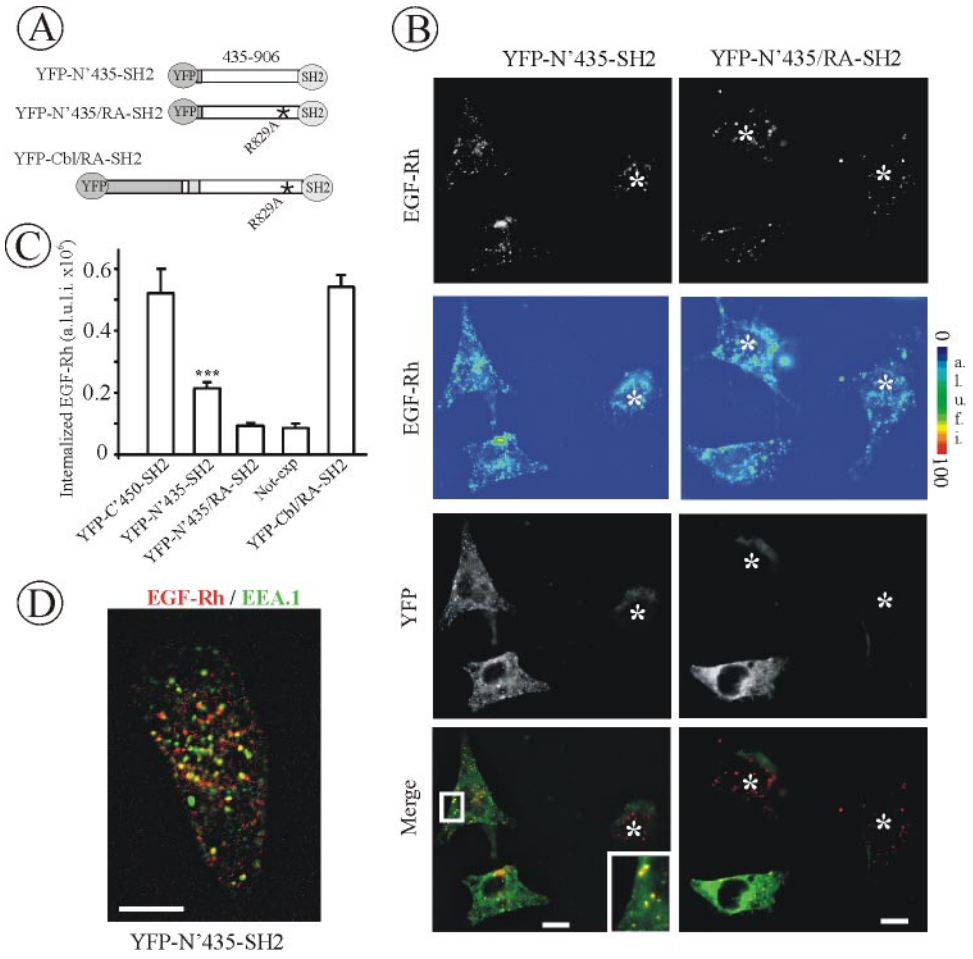


Figure 9. CIN85 interaction with Cbl is dispensable for EGFR internalization. (A) Schematic representation of Cbl-Grb2 chimeric proteins fused to YFP. The YFP-N′435-SH2 chimera consists of YFP, residues 435–906 of c-Cbl, and the SH2 domain of Grb2. YFP-N′435/RA-SH2 is a mutant YFP-N′435-SH2 chimera with the substitution of Arg829 (essential for CIN85 binding) to alanine in the c-Cbl fragment. YFP-Cbl/RA-SH2 contains YFP, full-length c-Cbl with the R829A mutation, and the SH2 domain of Grb2. (B) HeLa cells were transfected with Grb2 siRNA and either YFP-N′435-SH2 or YFP-N′435/RA-SH2. Cells were incubated with EGF-Rh, fixed, and imaged as described in Figure 3. YFP and Cy3 images representing individual optical sections were merged (Merge). EGF-Rh images also are displayed in quantitative pseudocolor using the same scale for both images. A.l.u.f.i., arbitrary linear units of fluorescence intensity. White asterisks show positions of Grb2-depleted cells that do not express YFP fusion proteins (nonexp). Inset represents high-magnification image of the region shown by white rectangles and demonstrate localization of the chimeric protein in endosomes containing EGF-Rh. Bars, 10 μm. (C) Quantification of the amount of internalized EGF-Rh from experiments similar to those presented in Figures 7B and 9B. In experiments with all YFP-tagged proteins, the values of integrated intensity did not depend on the expression levels of these fusion proteins. The error bars represent SEs ($n = 20$). ***$p < 0.0001$ compared with cells depleted of Grb2 and not expressing YFP-tagged constructs (nonexp). (D) HeLa cells transfected with Grb2 siRNA and YFP-N′435-SH2 were incubated with EGF-Rh for 15 min at 37°C, fixed, and processed for immunostaining with EEA.1 antibody. The z-stack of optical sections was acquired through the Cy3 (red), Cy5 (green) and YFP (our unpublished data) channels and deconvoluted. Cy5 and Cy3 images representing individual optical sections were merged (Merge). Yellow signifies colocalization of rhodamine and Cy5 labels. Bar, 10 μm.

activities of various proteins that were tagged with green fluorescent protein or its variants and transiently or stably expressed in mammalian cells. However, the caveats of these types of experiments are the presence of an endogenous protein and overexpression of tagged proteins, which often leads to their mistargeting and altered activities. Our work offers a simple approach of creating a human cell model system that allows visualization and analysis of the behavior and function of fluorescently labeled proteins expressed at physiological levels without the interference of the endogenous protein. Even initial comparison of Grb2 trafficking in this system and in cells overexpressing Grb2 revealed significant differences. Whereas overexpressed Grb2-YFP was robustly recruited to the plasma membrane ruffles upon EGF treatment (Sorkin *et al*., 2000), no ruffle targeting of Grb2-YFP was observed in cells expressing physiological levels of Grb2-YFP (Figure 2B).

The knockdown and rescue approach also has enabled us to dissect the mechanisms by which Grb2 interaction with Cbl regulates EGFR endocytosis. Because of the controversial literature on the function of Cbl in EGFR internalization, we initially attempted to use siRNA approach to directly test for the role of Cbl proteins in EGFR endocytosis. After examination of 15 different siRNA duplexes targeted to Cbl-b and c-Cbl, including those recently published (Mitra *et al*., 2004), we have not been able to identify siRNAs that produced >90% decrease in the combined immunoreactivity of c-Cbl and Cbl-b. Under these conditions EGFR inter-

nalization rates were reduced by 50%. It is possible that the residual pool of c-Cbl/Cbl-b proteins is sufficient to support endocytosis of a small number of activated EGFR under conditions of our internalization assay. It is also possible that Cbl-3 (Cbl-c), which can interact with Grb2 (Courbard *et al*., 2002), promotes EGFR ubiquitylation and endocytosis in the absence of c-Cbl and Cbl-b. Similarly, residual Cbl-b and Cbl-3 in c-Cbl−/− mouse embryonic fibroblasts could adequately support EGFR endocytosis in these cells (Duan *et al*., 2003). Therefore, the functional redundancy in Cbl family prompted us to examine the role of Grb2–Cbl interactions by using Grb2-Cbl chimeric proteins in the context of knockdown and rescue assay.

The experiments with Cbl-Grb2 chimeras demonstrated that the recruitment of Cbl to Grb2 binding sites of the EGFR is sufficient to completely restore receptor internalization in the absence of Grb2 (Figures 3 and 4). This finding suggests that Grb2 interactions with dynamin, SOS, or other SH3 binding proteins, such as POB1 (Nakashima *et al*., 1999), are dispensable, if not important at all, for EGFR endocytosis. In fact, Grb2 SH2 chimeras with SOS1 or dynamin II that were constructed analogously to the Cbl-Grb2 chimeras did not rescue EGFR endocytosis in Grb2-depleted cells (our unpublished data). On the other hand, we were able to generate stable HeLa cell lines, in which Grb2 was down-regulated by shRNA, whereas the SOS-Grb2 chimera was stably expressed (our unpublished data). The growth compensation effect of the SOS-Grb2 chimera is reminiscent to the mouse knockin studies, in which a similar Grb2-SOS chimera rescued the embryonic lethality caused by Grb2 knockout (Cheng *et al*., 1998). In contrast, growth defects due to Grb2 depletion could not be compensated by expression of the YFP-Cbl-SH2 chimera (our unpublished data).

Our data emphasized the importance of indirect, Grb2-mediated binding of Cbl to the EGFR for the regulation of receptor internalization. Cbl contains a variant of SH2 domain. However, this SH2-like domain alone cannot bind phosphotyrosine-containing motifs but requires the cooperation of the entire amino-terminal TKB domain (Meng *et al*., 1999). The direct binding of Cbl TKB domain to phosphorylated Tyr1045 (pTyr1045) was demonstrated using in vitro competition experiments with v-Cbl, a truncated variant of c-Cbl consisting of the TKB domain and a pTyr1045-containing peptide (Levkowitz *et al*., 1999). However, the same peptide did not interfere with the interaction of the full-length c-Cbl and EGFR, suggesting the prevailing role of other Cbl-EGFR association determinants (Levkowitz *et al*., 1999). In our experiments overexpression of c-Cbl, capable of direct binding to the EGFR, did not lead to stable association of c-Cbl with EGFR in the absence of Grb2, as judged by the absence of coimmunoprecipitation, and did not functionally compensate for Grb2 knockdown (Figures 3, 5, and 6). Thus, we propose that within the receptor dimer or monomer, the association of the carboxy-terminal proline-rich sequences of Cbl with Grb2 bound to EGFR allows Cbl TKB domain to be stably associated with EGFR pTyr1045, whereas the affinity of the latter interaction is not sufficient to sustain the association without the cooperation of the SH3-mediated interaction. Our findings are consistent with the model of Cbl interaction with the c-Met receptor, although unlike EGFR, the site of direct Cbl interaction with c-Met is in the juxtamembrane portion of the receptor (Peschard *et al*., 2001). Whereas Grb2-mediated association is the main mechanism of Cbl recruitment to the EGFR and is sufficient for EGFR internalization, interaction of Cbl TKB domain with pTyr1045 is essential for the maximal ubiquitylation and lysosomal targeting of the receptor (Levkowitz *et al*., 1999; Waterman *et al*., 2002; Jiang *et al*., 2003; Jiang and Sorkin, 2003).

In light of this model, it is surprising that functional SH2-like domain of Cbl was necessary for the full internalization-rescue activity of Cbl amino terminus (Figures 4 and 6). These data disagree with the observations that overexpression of c-Cbl G306E mutant did not affect EGFR internalization (Jiang and Sorkin, 2003). This discrepancy emphasizes the importance of the knockdown and rescue approach as opposed to an approach of a plain overexpression of the dominant-negative mutants that often suffers from the insufficient levels of mutant expression and presence of endogenous proteins. Because Cbl binding site in the EGFR (Tyr1045) is not necessary for internalization (Jiang and Sorkin, 2003; Oksvold *et al*., 2003; Grovdal *et al*., 2004), it is possible that TKB domain of Cbl has a function(s) additional to interaction with phosphotyrosine residues, as suggested previously (Meng *et al*., 1999).

The amino-terminal domain of c-Cbl, containing functional RING domain, fully restored EGFR endocytosis when recruited to the receptor by the SH2 domain of Grb2 (Figure 7). Furthermore, other SH2 domains, such as PLCγ1 N-SH2, capable of high-affinity interaction with EGFR, could replace Grb2 SH2 domain in the context of the RING-containing chimeric proteins in their capacity to restore EGFR endocytosis in Grb2-depleted cells (Figure 8). Although c-Cbl is capable of interaction with PLCγ1 SH3 domain (Tvorogov and Carpenter, 2002; Choi *et al*., 2003), PLCγ1−/− fibroblasts display normal EGFR trafficking (Ji *et al*., 1998). This suggests that Grb2-mediated binding of Cbl to EGFR is the main physiological mechanism of Cbl recruitment to the EGFR; however, in the absence of Grb2, association of Cbl through PLCγ1and other proteins can be a compensatory mechanism. These data also suggest that sustained association with EGFR is critical for the rescue activity of the Cbl amino terminus. Because SH2 domains of PLCγ1 and other proteins, when overexpressed, can bind to various phosphotyrosines of EGFR (Soler *et al*., 1994b; Chattopadhyay *et al*., 1999), it is difficult to determine whether a particular positioning of Cbl on the EGFR is most effective to support EGFR internalization.

Analysis of various mutant Cbl-Grb2 chimeras strongly indicated that the function of the RING domain is essential for the rescue effect of these chimeras on EGFR endocytosis (Figures 3–7). These data are consistent with the observations of the dominant-negative effects of overexpression of RING domain mutants of c-Cbl (Thien *et al*., 2001; Jiang and Sorkin, 2003) and Sprouty 2 that binds to the RING domain of Cbl (Stang *et al*., 2004) on EGFR internalization. However, the effect of Sprouty 2 cannot be unequivocally interpreted because Sprouty also directly binds to Grb2 and, when overexpressed, reduces Cbl binding to the EGFR.

The major and a well recognized function of the RING domain of Cbl is to recruit the ubiquitin ligase complex that mediates ubiquitylation of the EGFR. Alternatively, binding of proteins other than E2 ubiquitin ligases to the RING domain may link the EGFR–Grb2–Cbl complex to the clathrin coat. An example of RING-binding protein is Sprouty; however, this protein is thought to play a negative role in EGFR endocytosis (Wong *et al*., 2002; Stang *et al*., 2004). Two observations made in our study support the notion that the ubiquitylation function of Cbl is important for EGFR endocytosis mediated by Cbl-Grb2 chimeras. First, two different mutations in Cbl, both known to inhibit RING domain function through distinct mechanisms, have inactivated this chimera (Figures 5 and 7). One of the mutated residues, Trp408, directly participates in the interaction of the RING with the

E2 ligase (Zheng *et al*., 2000). Second, expression of YFP-Cbl-SH2 protein in Grb2-depleted cells restored EGFR ubiquitylation in a manner absolutely dependent on the intact RING domain (Figure 6). On the other hand, dramatic inhibition of EGFR ubiquitylation in EGFR Y1045F mutant did not result in significant inhibition of its internalization (Jiang and Sorkin, 2003; Grovdal *et al*., 2004). This suggests that if EGFR ubiquitylation is indeed involved in internalization, very small number of ubiquitination sites in the EGFR is sufficient to support this process.

Although our experiments established the essential role of the RING domain, they also confirmed that at least in HeLa cells, binding to CIN85 to the carboxy terminus of Cbl can function as an internalization mechanism, although this mechanism is significantly less effective than the RING-dependent mechanism. Because mutation of CIN85 binding site in full-length Cbl-Grb2 chimera did not affect the rescue capacity of this chimera, we suggest that CIN85-mediated mechanism is not normally involved in Grb2- and Cbl-mediated endocytosis unless the function of the RING domain is unavailable. This is consistent with our previous observations that overexpression of dominant-negative mutants of CIN85 did not have a specific inhibitory effect on EGFR internalization in HeLa cells (Jiang and Sorkin, 2003).

Assuming that EGFR ubiquitylation is indeed important for receptor endocytosis, it remains unknown how exactly this ubiquitylation participates in the clathrin-dependent process. Three coated pit proteins, epsin, Eps15, and Eps15R, contain ubiquitin-interacting domains. It is therefore tempting to suggest that ubiquitinated EGFR can bind to one of these proteins. For instance, in some cells Eps15 can be recruited to the plasma membrane in an EGF-dependent manner and positioned in the same location at the rim of coated pits as the EGFR–Grb2–Cbl complex (Stang *et al*., 2004). However, siRNA knockdown analysis of epsin 1, Eps15, and Eps15R did not provide support for the notion of the specific role of these proteins in clathrin-mediated internalization of EGFR (Huang *et al*., 2004). Mapping the ubiquitylation sites in the EGFR that are specifically responsible for the internalization step of endocytic trafficking of the receptor will be necessary to unequivocally establish the role of ubiquitylation and may allow identification of the binding partners of these ubiquitins in the clathrin-coated pits.

ACKNOWLEDGMENTS

We thank Dr. G. Carpenter for PLCγ1 DNA, Dr. E. Galperin for YFP-Cbl construct, and members of Sorkin laboratory for critical reading of the manuscript. This work was supported by grants from the National Cancer Institute and the American Cancer Society.

REFERENCES

Ahn, S., Kim, J., Lucaveche, C. L., Reedy, M. C., Luttrell, L. M., Lefkowitz, R. J., and Daaka, Y. (2002). Src-dependent tyrosine phosphorylation regulates dynamin self-assembly and ligand-induced endocytosis of the EGF receptor. J. Biol. Chem. *277*, 26642–26651.

Batzer, A. G., Rotin, D., Urena, J. M., Skolnik, E. Y., and Schlessinger, J. (1994). Hierarchy of binding sites for Grb2 and Shc on the EGF receptor. Mol. Cell. Biol. *14*, 5192–5201.

Chang, C.-P., *et al*. (1993). Ligand-induced internalization of the EGF receptor is mediated by multiple endocytic codes analogous to the tyrosine motif found in constitutively internalized receptors. J. Biol. Chem. *268*, 19312–19320.

Chattopadhyay, A., Vecchi, M., Ji, Q., Mernaugh, R., and Carpenter, G. (1999). The role of individual SH2 domains in mediating association of phospholipase C-gamma1 with the activated EGF receptor. J. Biol. Chem. *274*, 26091–26097.

Chen, W. S., Lazar, C. S., Lund, K. A., Welsh, J. B., Chang, C. P., Walton, G. M., Der, C. J., Wiley, H. S., Gill, G. N., and Rosenfeld, M. G. (1989). Functional independence of the epidermal growth factor receptor from a domain required for ligand-induced internalization and calcium regulation. Cell *59*, 33–43.

Cheng, A. M., *et al*. (1998). Mammalian Grb2 regulates multiple steps in embryonic development and malignant transformation. Cell *95*, 793–803.

Choi, J. H., Bae, S. S., Park, J. B., Ha, S. H., Song, H., Kim, J. H., Cocco, L., Ryu, S. H., and Suh, P. G. (2003). Cbl competitively inhibits EGF-induced activation of phospholipase C-gamma1. Mol. Cells *15*, 245–255.

Courbard, J. R., Fiore, F., Adelaide, J., Borg, J. P., Birnbaum, D., and Ollendorff, V. (2002). Interaction between two ubiquitin-protein isopeptide ligases of different classes, CBLC and AIP4/ITCH. J. Biol. Chem. *277*, 45267–45275.

de Melker, A. A., van der Horst, G., Calafat, J., Jansen, H., and Borst, J. (2001). c-Cbl ubiquitinates the EGF receptor at the plasma membrane and remains receptor associated throughout the endocytic route. J. Cell Sci. *114*, 2167–2178.

Dikic, I., and Giordano, S. (2003). Negative receptor signalling. Curr. Opin. Cell Biol. *15*, 128–135.

Duan, L., *et al*. (2003). Cbl-mediated ubiquitinylation is required for lysosomal sorting of EGF receptor but is dispensable for endocytosis. J. Biol. Chem. *278*, 28950–28960.

Grovdal, L. M., Stang, E., Sorkin, A., and Madshus, I. H. (2004). Direct interaction of Cbl with pTyr 1045 of the EGF receptor (EGFR) is required to sort the EGFR to lysosomes for degradation. Exp. Cell Res. *300*, 388–395.

Huang, F., Jiang, X., and Sorkin, A. (2003). Tyrosine phosphorylation of the beta2 subunit of clathrin adaptor complex AP-2 reveals the role of a di-leucine motif in the EGF receptor trafficking. J. Biol. Chem. *278*, 43411–43417.

Huang, F., Khvorova, A., Marshall, W., and Sorkin, A. (2004). Analysis of clatyhrin-mediated internalizaiton of EGF receptor by RNA interference. J. Biol. Chem. *In press*.

Ji, Q. S., Ermini, S., Baulida, J., Sun, F. L., and Carpenter, G. (1998). EGF signaling and mitogenesis in Plcg1 null mouse embryonic fibroblasts. Mol. Biol. Cell *9*, 749–757.

Jiang, X., Huang, F., Marusyk, A., and Sorkin, A. (2003). Grb2 Regulates Internalization of EGF Receptors through Clathrin-coated Pits. Mol. Biol. Cell *14*, 858–870.

Jiang, X., and Sorkin, A. (2003). EGF receptor internalization through clathrin-coated pits requires Cbl RING finger and proline-rich domains but not receptor polyubiquitylation. Traffic *4*, 529–543.

Johannessen, L. E., Ringerike, T., Molnes, J., and Madshus, I. H. (2000). EGF receptor efficiently activates mitogen-activated protein kinase in HeLa cells and hep2 cells conditionally defective in clathrin-dependent endocytosis. Exp. Cell Res. *260*, 136–145.

Kowanetz, K., Szymkiewicz, I., Haglund, K., Kowanetz, M., Husnjak, K., Taylor, J. D., Soubeyran, P., Engstrom, U., Ladbury, J. E., and Dikic, I. (2003). Identification of a novel proline-arginine motif involved in CIN85-dependent clustering of Cbl and down-regulation of EGF receptors. J. Biol. Chem. *278*, 39735–39746.

Kranenburg, O., Verlaan, I., and Moolenaar, W. H. (1999). Gi-mediated tyrosine phosphorylation of Grb2 (growth-factor-receptor-bound protein 2)-bound dynamin-II by lysophosphatidic acid. Biochem. J. *339*, 11–14.

Levkowitz, G., *et al*. (1999). Ubiquitin ligase activity and tyrosine phosphorylation underlie suppression of growth factor signaling by c-Cbl/Sli-1 [In Process Citation]. Mol. Cell *4*, 1029–1040.

Li, N., Batzer, A., Daly, R., Yajnik, V., Skolnik, E., Chardin, P., Bar-Sagi, D., Margolis, B., and Schlessinger, J. (1993). Guanine-nucleotide-releasing factor hSos1 binds to Grb2 and links receptor tyrosine kinases to Ras signalling [see comments]. Nature *363*, 85–88.

Marmor, M. D., and Yarden, Y. (2004). Role of protein ubiquitylation in regulating endocytosis of receptor tyrosine kinases. Oncogene *23*, 2057–2070.

Matsuda, M., Paterson, H. F., Rodriguez, R., Fensome, A. C., Ellis, M. V., Swann, K., and Katan, M. (2001). Real time fluorescence imaging of PLC gamma translocation and its interaction with the EGF receptor. J. Cell Biol. *153*, 599–612.

Meisner, H., Conway, B. R., Hartley, D., and Czech, M. P. (1995). Interactions of Cbl with Grb2 and phosphatidylinositol 3′-kinase in activated Jurkat cells. Mol. Cell. Biol. *15*, 3571–3578.

Meng, W., Sawasdikosol, S., Burakoff, S. J., and Eck, M. J. (1999). Structure of the amino-terminal domain of Cbl complexed to its binding site on ZAP-70 kinase. Nature *398*, 84–90.

Mitra, P., Zheng, X., and Czech, M. P. (2004). RNAi-based analysis of CAP, Cbl, and CrkII function in the regulation of GLUT4 by insulin. J. Biol. Chem. *279*, 37431–37435.

Motley, A., Bright, N. A., Seaman, M. N., and Robinson, M. S. (2003). Clathrin-mediated endocytosis in AP-2-depleted cells. J. Cell Biol. *162*, 909–918.

Nakashima, S., Morinaka, K., Koyama, S., Ikeda, M., Kishida, M., Okawa, K., Iwamatsu, A., Kishida, S., and Kikuchi, A. (1999). Small G protein Ral and its downstream molecules regulate endocytosis of EGF and insulin receptors. EMBO J. *18*, 3629–3642.

Oksvold, M. P., Thien, C. B., Widerberg, J., Chantry, A., Huitfeldt, H. S., and Langdon, W. Y. (2003). Serine mutations that abrogate ligand-induced ubiquitination and internalization of the EGF receptor do not affect c-Cbl association with the receptor. Oncogene. *22*, 8509–8518.

Okutani, T., Okabayashi, Y., Kido, Y., Sugimoto, Y., Sakaguchi, K., Matuoka, K., Takenawa, T., and Kasuga, M. (1994). Grb2/Ash binds directly to tyrosines 1068 and 1086 and indirectly to tyrosine 1148 of activated human EGF receptors in intact cells. J. Biol. Chem. *269*, 31310–31314.

Pearson, G., Robinson, F., Beers Gibson, T., Xu, B. E., Karandikar, M., Berman, K., and Cobb, M. H. (2001). Mitogen-activated protein (MAP) kinase pathways: regulation and physiological functions. Endocr. Rev. *22*, 153–183.

Peschard, P., Fournier, T. M., Lamorte, L., Naujokas, M. A., Band, H., Langdon, W. Y., and Park, M. (2001). Mutation of the c-Cbl TKB domain binding site on the Met receptor tyrosine kinase converts it into a transforming protein. Mol. Cell *8*, 995–1004.

Rozakis-Adcock, M., Fernley, R., Wade, J., Pawson, T., and Bowtell, D. (1993). The SH2 and SH3 domains of mammalian Grb2 couple the EGF receptor to the Ras activator mSos1 [see comments]. Nature *363*, 83–85.

She, H. Y., Rockow, S., Tang, J., Nishimura, R., Skolnik, E. Y., Chen, M., Margolis, B., and Li, W. (1997). Wiskott-Aldrich syndrome protein is associated with the adapter protein Grb2 and the EGF receptor in living cells. Mol. Biol. Cell *8*, 1709–1721.

Soler, C., Alvarez, C. V., Beguinot, L., and Carpenter, G. (1994a). Potent SHC tyrosine phosphorylation by EGF at low receptor density or in the absence of receptor autophosphorylation sites. Oncogene *9*, 2207–2215.

Soler, C., Beguinot, L., and Carpenter, G. (1994b). Individual EGF receptor autophosphorylation sites do not stringently define association motifs for several SH2-containing proteins. J. Biol. Chem. *269*, 12320–12324.

Sorkin, A., McClure, M., Huang, F., and Carter, R. (2000). Interaction of EGF receptor and grb2 in living cells visualized by fluorescence resonance energy transfer (FRET) microscopy. Curr. Biol. *10*, 1395–1398.

Sorkina, T., Huang, F., Beguinot, L., and Sorkin, A. (2002). Effect of tyrosine kinase inhibitors on clathrin-coated pit recruitment and internalization of EGF receptor. J. Biol. Chem. *277*, 27433–27441.

Soubeyran, P., Kowanetz, K., Szymkiewicz, I., Langdon, W. Y., and Dikic, I. (2002). Cbl-CIN85-endophilin complex mediates ligand-induced downregulation of EGF receptors. Nature *416*, 183–187.

Stang, E., Blystad, F. D., Kazazic, M., Bertelsen, V., Brodahl, T., Raiborg, C., Stenmark, H., and Madshus, I. H. (2004). Cbl-dependent ubiquitination is required for progression of EGF receptors into clathrin-coated pits. Mol. Biol. Cell *15*, 3591–3604.

Thien, C. B., and Langdon, W. Y. (2001). Cbl: many adaptations to regulate protein tyrosine kinases. Nat. Rev. Mol. Cell. Biol. *2*, 294–307.

Thien, C. B., Walker, F., and Langdon, W. Y. (2001). RING finger mutations that abolish c-Cbl-directed polyubiquitination and downregulation of the epidermal growth factor receptor are insufficient for cell transformation. Mol. Cell *7*, 355–365.

Tvorogov, D., and Carpenter, G. (2002). Epidermal growth factor-dependent association of phospholipase C-gamma1 with c-Cbl. Exp. Cell Res. *277*, 86–94.

Wang, X. J., Liao, H. J., Chattopadhyay, A., and Carpenter, G. (2001). epidermal growth factor-dependent translocation of green fluorescent protein-tagged PLC-gamma1 to the plasma membrane and endosomes. Exp. Cell Res. *267*, 28–36.

Wang, Y., and Wang, Z. (2003). Regulation of epidermal growth factor-induced phospholipase C-gamma1 translocation and activation by its SH2 and PH domains. Traffic *4*, 618–630.

Wang, Z., and Moran, M. F. (1996). Requirement for the adapter protein GRB2 in epidermal growth factor receptor endocytosis. Science *272*, 1935–1939.

Waterman, H., Katz, M., Rubin, C., Shtiegman, K., Lavi, S., Elson, A., Jovin, T., and Yarden, Y. (2002). A mutant epidermal growth factor-receptor defective in ubiquitylation and endocytosis unveils a role for Grb2 in negative signaling. EMBO J. *21*, 303–313.

Wong, E. S., Fong, C. W., Lim, J., Yusoff, P., Low, B. C., Langdon, W. Y., and Guy, G. R. (2002). Sprouty2 attenuates epidermal growth factor receptor ubiquitylation and endocytosis, and consequently enhances Ras/ERK signalling. EMBO J. *21*, 4796–4808.

Yamazaki, T., Zaal, K., Hailey, D., Presley, J., Lippincott-Schwartz, J., and Samelson, L. E. (2002). Role of Grb2 in epidermal growth factor-stimulated EGFR internalization. J. Cell Sci. *115*, 1791–1802.

Zheng, N., Wang, P., Jeffrey, P. D., and Pavletich, N. P. (2000). Structure of a c-Cbl-UbcH7 complex: RING domain function in ubiquitin-protein ligases. Cell *102*, 533–539.

nature structural & molecular biology

Determinants of nucleosome positioning

Kevin Struhl[1] & Eran Segal[2,3]

Nucleosome positioning is critical for gene expression and most DNA-related processes. Here we review the dominant patterns of nucleosome positioning that have been observed and summarize the current understanding of their underlying determinants. The genome-wide pattern of nucleosome positioning is determined by the combination of DNA sequence, ATP-dependent nucleosome remodeling enzymes and transcription factors that include activators, components of the preinitiation complex and elongating RNA polymerase II. These determinants influence each other such that the resulting nucleosome positioning patterns are likely to differ among genes and among cells in a population, with consequent effects on gene expression.

Eukaryotic genomes are packaged into chromatin, whose basic repeating unit is a nucleosome that consists of a histone octamer wrapped around 147 base pairs (bp) of DNA[1]. Nucleosomes are arranged into regularly spaced arrays, with the length of the linker region between nucleosomes varying among species and cell types. Although initially nucleosomes were believed to provide a universal, nonspecific coating of genomic DNA, it has long been known that nucleosomes occupy favored positions throughout the genome. High-resolution, genome-wide analyses have revealed a common pattern: nucleosomes are depleted at many (but not all) enhancer, promoter and terminator regions, and they typically occupy preferred positions in genes and non-gene regions[2–9]. In yeast, the −1 and +1 nucleosomes flanking the promoter are located at highly preferred positions, and the extent of preferred nucleosome positioning gradually decreases from the 5′ to 3′ end of the coding region[4,7].

In this Review we will consider the mechanisms by which the genomic pattern of nucleosome positioning is achieved. Intense research efforts collectively have revealed that nucleosome positioning is not determined by any single factor but rather by the combined effects of several factors including DNA sequence, DNA-binding proteins, nucleosome remodelers and the RNA polymerase II (Pol II) transcription machinery. One aim of this Review is to resolve the apparent controversy (in large part generated by us and the late Jonathan Widom, to whom we dedicate this Review) about the role of DNA sequence in establishing the genomic pattern of nucleosome positioning. This controversy stems from differences in interpretation and imprecision of terms, as the experimental observations have been remarkably consistent and are uncontested. In this Review we will focus on nucleosome positioning in yeast, particularly the biochemical and genetic analyses that have been rarely performed in multicellular organisms. We suspect that many aspects of nucleosome positioning are conserved among eukaryotes, but one important difference is that the linker histone H1 in yeast is structurally atypical, and it is present at very low concentrations in comparison to core histones[10].

[1]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts, USA. [2]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel. [3]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel. Correspondence should be addressed to E.S. (eran.segal@weizmann.ac.il) or K.S. (kevin@hms.harvard.edu).

Received 3 December 2012; accepted 3 January 2013; published online 5 March 2013; doi:10.1038/nsmb.2506

Definition of nucleosome positioning and occupancy

We define the term 'nucleosome positioning' broadly to indicate where nucleosomes are located with respect to the genomic DNA sequence. Nucleosome positioning is a dynamic process, but sequencing-based mapping approaches identify the positions of individual nucleosomes in a single cell at a specific time. Nevertheless, nucleosome positions are typically discussed on a cell- and time-averaged basis. Nucleosome positioning can vary from perfect positioning, in which a nucleosome is located at a given 147-bp stretch in all DNA molecules in a cell population, to no positioning, in which nucleosomes are located at all possible genomic positions with equal frequency across a cell population (**Fig. 1a**).

Nucleosome positioning is related to, but is distinct from, 'nucleosome occupancy', which reflects the fraction of cells from the population in which a given region of DNA is occupied by a histone octamer (**Fig. 1b**). Although most genomic DNA is occupied by nucleosomes, many functional regions (promoters, enhancers and terminators) are depleted of nucleosomes (have low occupancy) and some regions are largely nucleosome-free.

Nucleosome occupancy and positioning are critical to biological outcomes[4,7,11,12], primarily because nucleosomes inhibit the access of other DNA-binding proteins to DNA. Highly accessible regions in genomes are identified by preferential restriction endonuclease cleavage[13] and DNase I hypersensitivity[14] analyses. Quantitative analysis of HinfI cleavage[15] or Leu3 binding[16] in yeast cells indicates that access of factors to target sites in nucleosome-depleted promoter regions is ~10–20-fold higher than to identical sequences that are associated with nucleosomes. More generally, as a consequence of their modes of binding, transcription factors differ a great deal with respect to how much their binding is inhibited by nucleosomes. A special class of 'pioneer' transcription factors (for example, FoxA and GATA) can bind their target sites in the context of nucleosomal DNA[17]. Such pioneer factors, via recruitment of nucleosome remodelers, can open up the local chromatin, thereby facilitating the binding of other transcription factors that otherwise would be blocked by nucleosomes. The TATA-binding protein, and hence the entire basic Pol II transcription machinery, virtually cannot bind

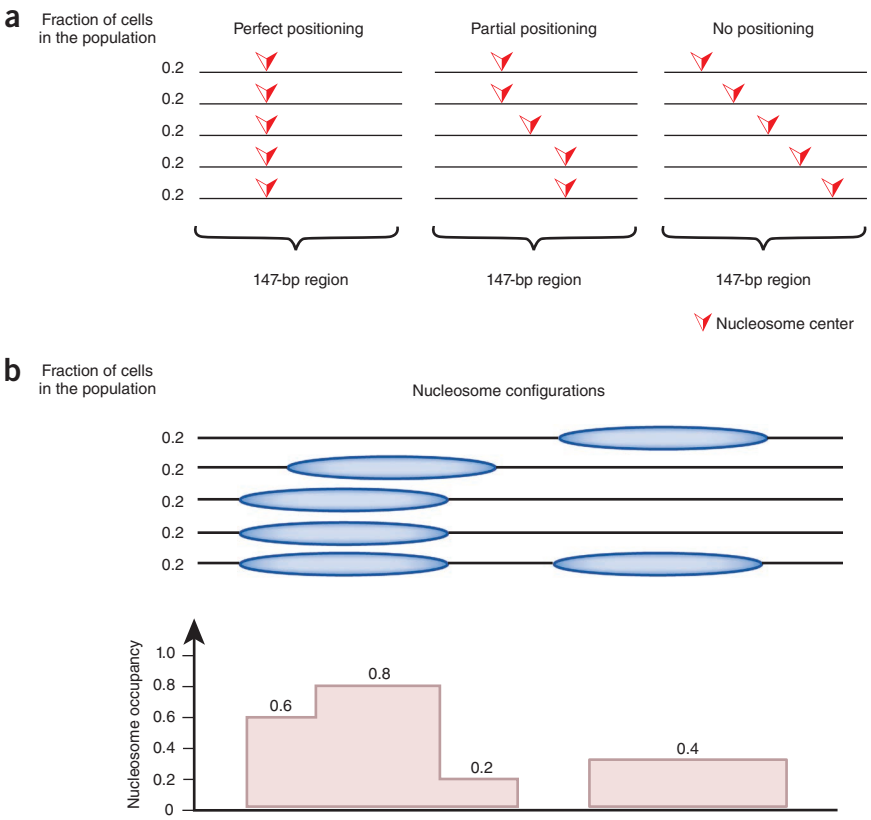


Figure 1 Illustration of the concepts of nucleosome positioning and nucleosome occupancy. (**a**) Nucleosome positioning along every base pair in the genome is defined as the fraction of cells from the population in which that base pair is at the center of a 147-bp nucleosome. For a given 147-bp region in a cell population, illustrated are perfect positioning, in which the center of the nucleosome is at the same base pair across all cells; partial positioning, in which there is a preference for some locations; and no positioning, in which all locations have equal probability. (**b**) Nucleosome occupancy along every base pair in the genome is defined as the fraction of cells from the population in which the base pair is occupied by any histone octamer. Illustrated are nucleosome locations across a cell population (top) and the resulting nucleosome occupancy per base pair (bottom).

nucleosomal DNA and hence requires a nucleosome-free region to bind core promoters and initiate transcription[18].

Nucleosome positioning is strongly affected by DNA sequence

The debate about the role of DNA sequence in nucleosome positioning *in vivo* revolves around the intrinsic sequence preferences of the histone octamer in the absence of any other component. The affinity of histone octamers for a given 147-bp sequence varies over more than three orders of magnitude[19]. As such, histone octamers exhibit considerable DNA sequence specificity, albeit lower than that of a classical specific DNA-binding protein. However, unlike DNA-binding proteins that achieve specificity by virtue of direct and strong interactions between a few base pairs and amino acids, the specificity of nucleosome formation largely reflects the overall ability of a given 147-bp sequence to bend around the histone octamer[20]. For optimal nucleosome formation, more bendable sequences are in contact with the histones, and less bendable sequences are solvent-exposed.

Two major sequence determinants affect bending and hence nucleosome formation. First, dinucleotides vary considerably with respect to their bending properties. Optimal nucleosome formation occurs when bendable dinucleotides (AT and TA) occur on the face of the helical repeat (every 10 bp) that directly interacts with histones[21,22]. Recent mapping of nucleosomes in the yeast *Saccharomyces cerevisiae*—using a H4-S47C–mediated cleavage technique that allows precise mapping of nucleosomes with respect to the DNA sequence—revealed 10-bp periodicity of bendable dinucleotides throughout nearly the entire 147-bp region[23]. The exact position of the histone octamer with respect to the ~10-bp helical repeat is termed 'rotational positioning', and thus, DNA sequence is a critical determinant of how nucleosomes are rotationally positioned (**Fig. 2**).

Second, the homopolymeric sequences poly(dA:dT) and poly(dG:dC) are intrinsically stiff (for different structural reasons) and are strongly inhibitory to nucleosome formation[24–27]. Notably, poly(dA:dT) tracts are abundant in eukaryotic genomes[28] and are particularly prevalent in promoters of certain organisms such as *S. cerevisiae*[29,30]. As we will discuss below, intrinsic properties of poly(dA:dT) are important for nucleosome depletion, promoter accessibility and transcriptional activity[12,15,26].

By analogy to DNA-binding proteins, the above DNA sequence preferences for nucleosomes can be expressed as a position-weight matrix that can generate a nucleosome score for any 147-bp region of DNA[22]. These nucleosome scores correspond to relative affinities of 147-bp sequences for histone octamers; the differences in affinities must affect nucleosome positioning *in vivo*. In particular, as dinucleotide preferences vary along the 147-bp region of the nucleosome, most DNA sequences will have considerably different nucleosome

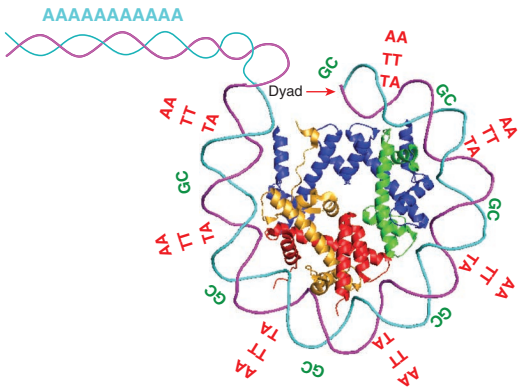


Figure 2 Nucleosome sequence preferences. Within the 147 bp that are wrapped around the histone octamer, there is a preference for distinctive dinucleotides that recur periodically at the DNA helical repeat and are known to facilitate the sharp bending of DNA around the nucleosome. These include ~10-bp periodic AA, TT or TA dinucleotides that oscillate in phase with each other and out of phase with ~10-bp periodic GC dinucleotides. The linker regions exhibit a strong preference for sequences that resist DNA bending and thus disfavor nucleosome formation. Among these, poly(dA:dT) tracts and their variants are most dominant and highly enriched in eukaryotic promoters.

scores at neighboring positions in an individual helical repeat. As a consequence, there will usually be a preferred position(s) in a helical repeat, and such rotational positioning is clearly observed in eukaryotic genomes. However, if one considers a favored rotational position, genomic locations that are 10 bp apart (1 helical turn) will have similar (although not identical) nucleosome formation scores, unless the nearby locations contain strongly inhibitory sequences[31]. Thus, in many genomic regions, DNA sequence alone is unlikely to strongly favor one nucleosome position over nearby positions that are rotationally equivalent.

Poly(dA:dT) tracts are important for nucleosome depletion

The role of DNA sequence in establishing nucleosome positions *in vivo* has been addressed in experiments in which nucleosomes are assembled *in vitro* by salt dialysis using purified histones and genomic DNA[32,33]. Under these conditions, many yeast promoter and terminator regions are depleted of nucleosomes, indicating that the DNA sequences intrinsically disfavor formation of nucleosomes. Consistent with this view, depletion of nucleosomes at promoters is observed under a wide variety of conditions *in vivo* and is unaffected by transcriptional activity[32,34]. In addition, when artificial chromosomes containing large genomic regions from heterologous yeast species are analyzed in *S. cerevisiae*, depletion of nucleosomes at promoters is maintained in a manner that depends on the number, length and fully homopolymeric nature of poly(dA:dT) sequences[35]. Thus, depletion of nucleosomes at most promoter sequences is strongly influenced by intrinsic properties of poly(dA:dT) sequences.

A variety of detailed functional analyses indicate that the intrinsic properties of poly(dA:dT) are important for nucleosome depletion, promoter accessibility and transcriptional activity[12,15,26,29,34–37]. In principle, poly(dA:dT)-mediated depletion could be due to a poly(dA:dT)-binding activator protein, but the one known factor with such DNA-binding specificity (Datin) actually inhibits transcription[15]. Consistent with an earlier study[15], expression measurements of a large-scale promoter library designed with systematic manipulations to the properties and spatial arrangement of poly (dA:dT) tracts showed that longer and more perfect tracts induce transcription regardless of the identity of the binding transcription factor[12]. This effect of the tracts on transcription is mediated by reduced nucleosome occupancy and thus increased accessibility that these tracts confer on nearby promoter elements, such as transcription factor–binding sites.

Alteration of poly(dA:dT) tracts offers a general evolutionary mechanism, applicable to promoters regulated by different transcription factors, for tuning expression in a predictable manner, with resolution that can be even finer than that attained by altering transcription factor–binding sites[12]. Indeed, genomes have likely used these tracts to regulate expression, for example, to partly compensate for differences in gene copy number that exist among ribosomal protein genes in *S. cerevisiae*[38] and to alter expression of cellular respiration genes across yeast species[39]. Moreover, partitioning yeast genes into two categories on the basis of the extent to which their promoters favor nucleosome formation in a manner largely dependent on the occurrences of poly(dA:dT) tracts results in functionally distinct gene classes. Specifically, the class of genes whose promoters have a higher intrinsic preference for nucleosomes is enriched for stress-response genes, exhibits higher rates of histone turnover and transcriptional noise, and contains more targets of chromatin remodelers, consistent with an ongoing dynamic competition between nucleosome assembly and factor binding[30,40]. Thus, poly(dA:dT) tracts are important determinants of nucleosome depletion *in vivo* and have likely been used by organisms to obtain biological outcomes.

Yeast terminator regions are also depleted of nucleosomes both *in vivo* and *in vitro*. In contrast to the situation at promoters, depletion of nucleosomes at terminator regions is strongly correlated with the orientation of and distance to neighboring genes, and it is strongly affected by growth conditions and transcriptional elongation by Pol II (ref. 41). Thus, the contribution of DNA sequence to depletion at terminators requires additional examination.

A minority of *S. cerevisiae* promoters are depleted for nucleosomes *in vivo*, yet lack poly(dA:dT) sequences and are not depleted in nucleosome assembly experiments *in vitro*. More generally, depletion of nucleosomes at promoters is observed in many species in which poly(dA:dT) sequences are less prevalent than in *S. cerevisiae* or are rare[42–44]. At these promoters, depletion of nucleosomes is not determined by DNA sequence but is likely due to activator-mediated recruitment of nucleosome remodelers that evict histones in promoter regions.

Aspects of positioning not determined by DNA sequence

Although nucleosome depletion can be reconstituted at most promoter sequences in nucleosome assembly experiments with purified histones and DNA, other aspects of the *in vivo* pattern of nucleosome positioning cannot be reconstituted[32,33]. In particular, strong positioning of the +1 nucleosome is not observed, and there is no evidence for any favored position in this region under conditions in which histone concentrations are either limiting or saturating. These observations suggest that the DNA sequence is not the main determinant of the position of the +1 nucleosome. Furthermore, positions of the +2 nucleosome and further downstream nucleosomes are determined primarily by the length of the linker region and hence by the position of the +1 nucleosome, and they are also not reconstituted *in vitro*.

Theoretically, the combination of a boundary constraint and high nucleosome concentrations along the DNA could generate an ordered array that begins with the +1 nucleosome and decays with increasing distance downstream. This 'statistical positioning' phenomenon could permit nucleosome-depleted regions mediated by poly(dA:dT) tracts to act as barriers and indirectly contribute to the long-range positioning of nucleosomes that flank promoters. However, statistically positioned arrays flanking poly(dA:dT) tracts were not observed even under *in vitro* conditions in which arrays were clearly generated[33], providing evidence against this theoretical possibility.

Overall, positions of nucleosomes that are assembled *in vitro* resemble those observed *in vivo* beyond random expectation, indicating that DNA sequence does contribute significantly to nucleosome positioning. However, in many cases, other factors can override the sequence preferences for nucleosome formation. Such overriding is likely to occur in the many genomic regions in which the dynamic range of affinities of the histone octamer to the underlying sequence is narrower than the more than three orders of magnitude observed for specific DNA sequences.

Role of nucleosome remodelers in nucleosome positioning

Several aspects of the *in vivo* nucleosome positioning pattern can be reconstituted *in vitro* if a yeast crude extract and ATP are added to purified histones and DNA[45]. Specifically, the combination of ATP-dependent nucleosome remodelers in a cell-free extract enhances depletion of nucleosomes at promoters to the extent observed *in vivo*, and it also generates positioned nucleosomes flanking the nucleosome-depleted regions (that is, the +1 and −1 nucleosomes). In principle, remodeling enzymes may simply allow nucleosomes to sample

positions rapidly, resulting in a thermodynamic equilibrium that is not achieved when nucleosome assembly is performed by salt dialysis. However, nucleosome assembly reactions containing *Drosophila melanogaster* ACF[33] or the yeast RSC[46] do not generate the *in vivo* pattern, even though they efficiently mobilize nucleosomes. These observations suggest that nucleosome-remodeling enzymes do not simply facilitate the movement of nucleosomes to preferred intrinsic positions, but rather are critical in establishing the specificity of where nucleosomes are located.

The mechanism by which nucleosome remodelers override intrinsic DNA sequence preferences of histone octamers is unclear. However, RSC[47,48] and perhaps other nucleosome remodelers can bind specific DNA sequences, and they are likely to have other sequence preferences for positioning nucleosomes[49,50]. In addition, nucleosome remodelers may be influenced by the boundary of nucleosome-disfavoring sequences, such that in the course of moving nucleosomes kinetic effects override thermodynamic equilibrium[50]. In this regard, nucleosome remodelers may use nucleosome-depleted regions as a measuring device to position the +1 and −1 nucleosomes. Notably, multiple nucleosome remodelers in yeast cell-free extracts are necessary to reconstitute the genome-wide pattern observed *in vivo*. Reactions involving individual nucleosome remodelers reconstitute only part of the pattern, either being restricted to subsets of genes or generating less precise positioning[46].

Although yeast nucleosome remodeling activities in themselves can reconstitute some aspects of the *in vivo* pattern, there are two important aspects of the pattern that they do not faithfully generate. First, the precise locations of the +1 nucleosomes generated *in vitro* poorly match those *in vivo*. Second, the extent of positioning of more downstream nucleosomes (for example, +3 and beyond) is substantially less in the *in vitro* reactions than what is observed *in vivo*. As a consequence, the nucleosome arrays over coding regions are less pronounced and appear shorter. These observations indicate that nucleosome remodelers are necessary but not sufficient to establish the *in vivo* pattern of nucleosome positioning. As we will discuss below, aspects of Pol II transcription also have critical roles in establishing the *in vivo* pattern.

In addition to data from *in vitro* reconstitution experiments, there is considerable genetic evidence for the critical role of nucleosome remodelers in establishing nucleosome positioning *in vivo*. In *S. cerevisiae*, a strain lacking the Isw2 remodeler shows altered nucleosome positioning adjacent to the promoter at the interface of genic and intergenic sequences[51]. Specifically, Isw2 helps to position the +1 nucleosome onto unfavorable DNA near the promoter in a directional manner, and it suppresses antisense transcription[52]. In addition, the RSC remodeling complex also mobilizes nucleosomes onto unfavorable sequences in the vicinity of the promoter[53].

Role of nucleosome remodelers in nucleosome spacing

As nucleosomes are typically arranged in regularly spaced arrays with a relatively constant linker length between nucleosomes, nucleosome spacing is a key aspect of nucleosome positioning. Drastic alteration of nucleosome positioning patterns is observed in an *S. cerevisiae* strain lacking both the Isw1 and Chd1 remodelers[54] or in an *S. pombe* strain lacking two related CHD remodelers (Hrp1 and Hrp3)[55–57]. In each of these evolutionarily diverged yeast species, positioning of the +2 nucleosome is much lower than that observed in the wild-type strain, and positioning of the +3 nucleosome and more downstream nucleosomes is essentially lost. This dramatic alteration of nucleosome positioning in coding regions is not due to histone depletion but rather to a defect in nucleosome spacing. Thus, a combination of the Isw1 and Chd1 (or Hrp1 and Hrp3) nucleosome remodeling enzymes is required for the correct nucleosome spacing that is the basis of positioned nucleosome arrays. Notably, a small subset of nucleosomes in downstream portions of coding regions are correctly positioned, presumably because of intrinsic DNA sequence preferences that facilitate or override the effect of nucleosome remodelers[54]. Positions of the +1 and −1 nucleosomes are essentially unaffected by the combined loss of Isw1 and Chd1 function, indicating that their positioning occurs by a distinct mechanism that may involve Isw2.

Additional support for a role of nucleosome remodelers for nucleosome spacing and positioning comes from a functional evolutionary experiment[35] that is based on the observation that nucleosome spacing varies among yeast species[43,44]. When large genomic regions from a foreign yeast species are introduced into *S. cerevisiae*, the distance between nucleosomes is characteristic of *S. cerevisiae*, not of the donor yeast species[35]. As a consequence, the vast majority of nucleosomes on the foreign DNA are not located at the positions that occur when the same DNA is present in the endogenous organism. The change in spacing could be due to differences in histone concentrations between the species, as the number of nucleosomes on genomic DNA will necessarily affect the average spacing. However, depletion of histone H3 does not generally alter nucleosome spacing, although some positioned nucleosomes are preferentially lost or maintained[55,58,59]. Or this observation may be due to species-specific differences in the spacing properties of the remodelers themselves because nucleosome assembly *in vivo* requires remodelers and remodelers have specific nucleosome-spacing properties that are independent of histone concentration[60].

In *S. cerevisiae*, histone H1 does not noticeably affect nucleosome spacing because histone H1 protein amounts are much lower than those of the core histones[10]. However, histone H1 and its various subtypes have an important role in nucleosome spacing in multicellular organisms. Overexpression or depletion experiments *in vivo* indicate that histone H1 increases the spacing between adjacent nucleosomes[61–64], and differences in linker histone subtypes might underlie cell type–specific differences in nucleosome spacing. In addition to histone H1, the HMG14, 17 proteins may also have a role in nucleosome spacing[61].

Role of transcription factors and Pol II elongation

In addition to DNA sequence and ATP-dependent nucleosome remodelers, Pol II transcription also contributes to establishing the genomic pattern of nucleosome positioning. In this respect, nucleosome positioning, transcription and perhaps other DNA-based processes such as DNA replication should be viewed as processes that reciprocally affect each other. The effect of Pol II on nucleosome positioning is achieved via transcriptional activator proteins, general transcription factors that compose the preinitiation complex and the elongating Pol II machinery.

Nucleosome depletion can generate positioning *in vivo*. Transcriptional activators, via targeted recruitment of nucleosome remodelers, can generate nucleosome-depleted regions. Some activator proteins can bind nucleosomal DNA fairly well, whereas others rely on intrinsic histone-destabilizing sequences or cooperativity with other activators to access their sites. Under standard growth conditions, the Rap1, Abf1 and Reb1 activators are important for generating nucleosome-depleted regions at subsets of *S. cerevisiae* genes[32,47,53]. When foreign DNA is introduced into *S. cerevisiae*, nucleosome-depleted regions often occur in coding regions, unlike the case in the native organisms[35]. Such *de novo* nucleosome-depleted regions

presumably arise from the fortuitous binding of *S. cerevisiae* activators to evolutionarily irrelevant target sites. Notably, many of these fortuitous nucleosome-depleted regions are associated with a positioned nucleosome array that strongly resembles the standard nucleosome positioning pattern of endogenous *S. cerevisiae* genes. Thus, generation of a nucleosome-depleted region, even in the absence of intrinsic nucleosome-destabilizing sequences, is sufficient to generate the *in vivo* pattern of nucleosome positioning.

Basal transcription factors help position the +1 nucleosome. The precise location of the +1 nucleosome is a critical determinant of the nucleosome positioning pattern because nucleosome spacing constraints are major determinants of the downstream nucleosome positions. The inability of the collection of nucleosome remodelers in crude extracts to accurately reconstitute the *in vivo* positions of +1 and −1 nucleosomes suggests that some other factor(s) is involved. The relationship between the position of the +1 nucleosome and the transcription start site (TSS) suggests that the general transcription factors have a role[33]. Furthermore, locations of the +1 nucleosome and TSS shift in concert when foreign yeast DNA is introduced into *S. cerevisiae*, such that TSS on the foreign DNA shifts to an *S. cerevisiae*–like location[35]. Given the strong *in vivo* positioning of both the preinitiation complex and the +1 nucleosome, a spacing relationship between these two entities requires that at least one of these be anchored to a specific location, thereby permitting a defined location for the second entity. Although the limited sequence specificity of nucleosome remodelers makes it unlikely that they can provide such an anchor, preinitiation complexes bound at core promoters may be sufficient, with the location of the TATA-binding protein bound to the TATA element or TATA-related sequence being the major determinant of the anchor point[65].

These considerations strongly suggest that the preinitiation complex has a role in fine-tuning the position of the +1 nucleosome. One speculative possibility is that a component(s) of the preinitiation complex is important for transiently recruiting the Isw2 and/or RSC complex that overrides inherent DNA sequence preferences to precisely position the +1 nucleosome. In addition, for organisms in which Pol II is often paused just downstream of the promoter[66], there is a strong distance relationship between the presence of paused Pol II and the NELF pausing factor and the position of the +1 nucleosome[67].

Pol II elongation and generating nucleosome arrays. Nucleosome arrays emanating from promoter regions occur unidirectionally in the transcribed direction, even though the +1 and −1 nucleosomes are well positioned. In addition, the decay of nucleosome positioning toward the center of genes displays a 5′-3′ asymmetry[68]. These initial observations suggested that the elongating Pol II machinery is important in establishing the pattern of nucleosome positioning. In accord with this idea, the Chd1 and Isw1 nucleosome remodelers that are critical for nucleosome positioning in coding regions have genome-wide association patterns that strongly resemble that of elongating Pol II. Conversely, nucleosome assembly in transcriptionally incompetent, cell-free extracts poorly recapitulates positioned nucleosome arrays in the downstream portions of coding regions, even though the +1 and −1 nucleosomes are strongly positioned. Lastly, the length of nucleosome arrays emanating from fortuitous nucleosome-depleted regions in foreign yeast DNA that act as functional promoters is strongly correlated both in direction and length with the mRNA[35].

These observations suggest that Pol II elongation strongly affects nucleosome positioning. In particular, the unidirectionality of nucleosome arrays can be easily explained by the unidirectionality of transcription, whereas it is unclear how such unidirectionality could be imposed only by nucleosome-depleted regions and nucleosome remodelers. Furthermore, the inefficiency of yeast cell extracts to reconstitute downstream nucleosome positions in the coding region suggests that recruitment of the Chd1 and Isw1 remodeling enzymes is not the sole mechanism by which elongating Pol II affects nucleosome positioning. Nevertheless, the mechanistic connection between Pol II elongation and nucleosome arrays in coding regions remains to be established.

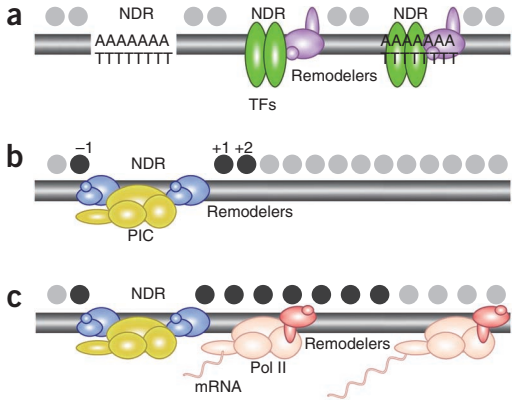


Figure 3 Determinants of nucleosome positioning. (**a**) Nucleosome-depleted regions (NDRs) are generated either by poly(dA:dT) tracts and/or by transcription factors and their recruited nucleosome remodeling complexes. Gray circles indicate nucleosomes. (**b**) Nucleosomes located at highly preferred positions (black circles) flanking the NDR are generated by nucleosome-remodeling complexes (for example, Isw2 and RSC, likely in a transcription-independent manner), and fine-tuned by the Pol II preinitiation complex (PIC) and associated factors. (**c**) Positioning of the more downstream nucleosomes depends on transcriptional elongation, and the recruitment of nucleosome-remodeling activities (for example, Chd1 and Isw1) and histone chaperones by the elongating Pol II machinery. This figure, which was modified from ref. 36, does not include DNA sequence determinants of rotational positioning (**Fig. 2**) or contributions to nucleosome spacing by histone H1.

Summary

The genetic, biochemical and informatics analyses performed in many laboratories together demonstrate that the genome-wide pattern of nucleosome positioning is determined by the combination of DNA sequence, nucleosome remodelers and transcription factors including activators, components of the preinitiation complex and elongating Pol II (**Fig. 3**). Although each of these components has discernible effects in isolation, they also reciprocally affect each other and hence affect the nucleosome positioning pattern in potentially complex ways. The DNA sequence is critical for rotational positioning along the DNA helix, and it also is an important determinant for nucleosome occupancy. In particular, poly(dA:dT) and poly(dG:dC) tracts are intrinsically inhibitory to nucleosome formation, whereas non-homopolymeric (G+C)-rich regions favor nucleosome formation. DNA sequence also contributes to the nucleosome positioning pattern, but several aspects of the *in vivo* pattern cannot be accounted for by intrinsic histone-DNA interactions.

As demonstrated by cross-species experiments, nucleosome-depleted regions are largely sufficient to generate the standard pattern of ordered nucleosome arrays[35]. At native promoters, nucleosome-depleted regions can be generated intrinsically via poly(dA:dT) tracts and/or activator-mediated recruitment of nucleosome remodelers that evict histones in the vicinity of the activator-binding site. The use of poly(dA:dT) sequences varies considerably among organisms;

they are very common in *S. cerevisiae* but not in the fission yeast *S. pombe* and in multicellular organisms[42–44].

In addition to their role in creating nucleosome-depleted regions, ATP-dependent remodelers are important in nucleosome positioning. In a manner independent of transcription, they can assemble nucleosomes flanking the depleted region to generate the +1 and −1 nucleosomes. However, these remodeling enzymes cannot accurately position the +1 nucleosome to the *in vivo* location nor can they generate properly positioned nucleosomes at more downstream locations. Precise positioning of the +1 nucleosome is strongly influenced by the location of the preinitiation complex, although the mechanistic basis for this spacing relationship is poorly understood. Lastly, generation of positioned nucleosome arrays throughout the coding region is coupled to Pol II elongation, and it involves Pol II–dependent recruitment of nucleosome remodelers and perhaps some other aspect of Pol II elongation. The putative roles of Isw2 and RSC for positioning the +1 nucleosome and of Isw1 and Chd1 for transcription-coupled positioning and spacing of nucleosomes in the coding region are strongly supported by genome-wide mapping of nucleosome remodelers on positioned mononucleosomes[69].

In some situations where a gene is activated, it is likely that establishment of the nucleosome pattern occurs stepwise: activator-mediated generation of a nucleosome-depleted region, positioning of the +1 and −1 nucleosomes, and elongation-coupled assembly of positioned and properly spaced nucleosome arrays over the coding region. However, such a stepwise process is unnecessary, and probably irrelevant, for steady-state maintenance of the nucleosome positioning pattern. Instead, the combined effects of DNA sequence, nucleosome remodelers and transcription factors make independent contributions to the pattern.

Lastly, it is important to mention that the nucleosome positioning pattern described here is gene-averaged and hence represents a typical pattern. Thus, although the positioning mechanisms discussed here apply to all genes, the precise pattern at individual genes may differ from the gene-averaged pattern. Such variations will depend on the DNA sequences, the nucleosome remodelers that act at the gene and the complexity of transcription units within the gene (for example, antisense or unstable transcripts). The relative contributions of these mechanisms may differ among genes, which is likely to result in differential regulation of these genes. Furthermore, even when the typical pattern predominates at a given gene, it is highly likely that the pattern is not present in all cells. For this reason, differences in the nucleosome positioning pattern among cells in a population are likely to confer distinct transcriptional properties in these cells. Nevertheless, although several questions remain open, we believe that there is now a good understanding of how many aspects of the nucleosome positioning patterns are generated *in vivo*.

ACKNOWLEDGMENTS

This Review is dedicated to the memory of Jon Widom, our dear friend and colleague, who is sorely missed by the authors and the scientific community. We thank D. Schneider for encouraging us to write this Review, and Z. Moqtaderi, T. Raveh-Sadka, M. Levo, N. Kaplan and Y. Field for comments on the manuscript. This work was supported by grant GM 30186 from the US National Institutes of Health to K.S. and grants from the European Research Council and the US National Institutes of Health to E.S.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/nsmb.2506.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Richmond, T.J. & Davey, C.A. The structure of DNA in the nucleosome core. *Nature* **423**, 145–150 (2003).
2. Yuan, G.-C. *et al.* Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science* **309**, 626–630 (2005).
3. Lee, W. *et al.* A high-resolution atlas of nucleosome occupancy in yeast. *Nat. Genet.* **39**, 1235–1244 (2007).
4. Mavrich, T.N. *et al.* A barrier nucleosome model for statistical positioning of nucleosome throughout the yeast genome. *Genome Res.* **18**, 1073–1083 (2008).
5. Mavrich, T.N. *et al.* Nucleosome organization in the *Drosophila* genome. *Nature* **453**, 358–362 (2008).
6. Schones, D.E. *et al.* Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132**, 887–898 (2008).
7. Shivaswamy, S. *et al.* Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biol.* **6**, e65 (2008).
8. Valouev, A. *et al.* A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res.* **18**, 1051–1063 (2008).
9. Valouev, A. *et al.* Determinants of nucleosome organization in primary human cells. *Nature* **474**, 516–520 (2011).
10. Freidkin, I. & Katcoff, D.J. Specific distribution of the *Saccharomyces cerevisiae* linker histone HHO1 in the chromatin. *Nucleic Acids Res.* **29**, 4043–4051 (2001).
11. Lam, F.H., Steger, D.J. & O'Shea, E.K. Chromatin decouples promoter threshold from dynamic range. *Nature* **453**, 246–250 (2008).
12. Raveh-Sadka, T. *et al.* Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. *Nat. Genet.* **44**, 743–750 (2012).
13. Varshavsky, A.J., Sundin, O. & Bohn, M. A stretch of "late" SV40 viral DNA about 400 bp long which induces the origin of replication is specifically exposed in SV40 minichromosomes. *Cell* **16**, 453–466 (1979).
14. Wu, C. The 5′ ends of *Drosophila* heat shock genes in chromatin are hypersensitive to DNase I. *Nature* **286**, 854–860 (1980).
15. Iyer, V. & Struhl, K. Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic structure. *EMBO J.* **14**, 2570–2579 (1995).
16. Liu, X., Lee, C.K., Granek, J.A., Clarke, N.D. & Lieb, J.D. Whole-genome comparison of Leu3 binding *in vitro* and *in vivo* reveals the importance of nucleosome occupancy in target site selection. *Genome Res.* **16**, 1517–1528 (2006).
17. Zaret, K.S. & Carroll, J.S. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev.* **25**, 2227–2241 (2011).
18. Workman, J.L. & Kingston, R.E. Alteration of nucleosome structure as a mechanism of transcriptional regulation. *Annu. Rev. Biochem.* **67**, 545–579 (1998).
19. Thastrom, A. *et al.* Sequence motifs and free energies of selected natural and non-natural nucleosome positioning DNA sequences. *J. Mol. Biol.* **288**, 213–229 (1999).
20. Drew, H.R. & Travers, A.A. DNA bending and its relation to nucleosome positioning. *J. Mol. Biol.* **186**, 773–790 (1985).
21. Satchwell, S.C., Drew, H.R. & Travers, A.A. Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.* **191**, 659–675 (1986).
22. Segal, E. *et al.* A genomic code for nucleosome positioning. *Nature* **442**, 772–778 (2006).
23. Brogaard, K.R., Xi, L., Wang, J.P. & Widom, J. A map of nucleosome positions in yeast at base-pair resolution. *Nature* **486**, 496–501 (2012).
24. Nelson, H.C.M., Finch, J.T., Luisi, B.F. & Klug, A. The structure of an oligo(dA)·oligo(dT) tract and its biological implications. *Nature* **330**, 221–226 (1987).
25. Suter, B., Schnappauf, G. & Thoma, F. Poly(dA-dT) sequences exist as rigid DNA structures in nucleosome-free yeast promoters *in vivo*. *Nucleic Acids Res.* **28**, 4083–4089 (2000).
26. Segal, E. & Widom, J. Poly(dA:dT) tracts: major determinants of nucleosome organization. *Curr. Opin. Struct. Biol.* **19**, 65–71 (2009).
27. McCall, M., Brown, T. & Kennard, O. The crystal structure of d(G-G-G-G-C-C-C-C): a model for poly(dG)·poly(dC). *J. Mol. Biol.* **183**, 385–396 (1985).
28. Dechering, K.J., Cuelenaere, K., Konings, R.N. & Leunissen, J.A. Distinct frequency-distributions of homopolymeric DNA tracts in different genomes. *Nucleic Acids Res.* **26**, 4056–4062 (1998).
29. Struhl, K. Naturally occurring poly(dA-dT) sequences are upstream promoter elements for constitutive transcription in yeast. *Proc. Natl. Acad. Sci. USA* **82**, 8419–8423 (1985).
30. Field, Y. *et al.* Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Comput. Biol.* **4**, e1000216 (2008).
31. Gaffney, D.J. *et al.* Controls of nucleosome positioning in the human genome. *PLoS Genet.* **8**, e1003036 (2012).
32. Kaplan, N. *et al.* The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**, 362–366 (2009).
33. Zhang, Y. *et al.* Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions *in vivo*. *Nat. Struct. Mol. Biol.* **16**, 847–852 (2009).
34. Sekinger, E.A., Moqtaderi, Z. & Struhl, K. Intrinsic histone-DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast. *Mol. Cell* **18**, 735–748 (2005).
35. Hughes, A., Jin, Y., Rando, O.J. & Struhl, K. A functional evolutionary approach to identify determinants of nucleosome positioning: a unifying model for establishing the genome-wide pattern. *Mol. Cell* **48**, 5–15 (2012).

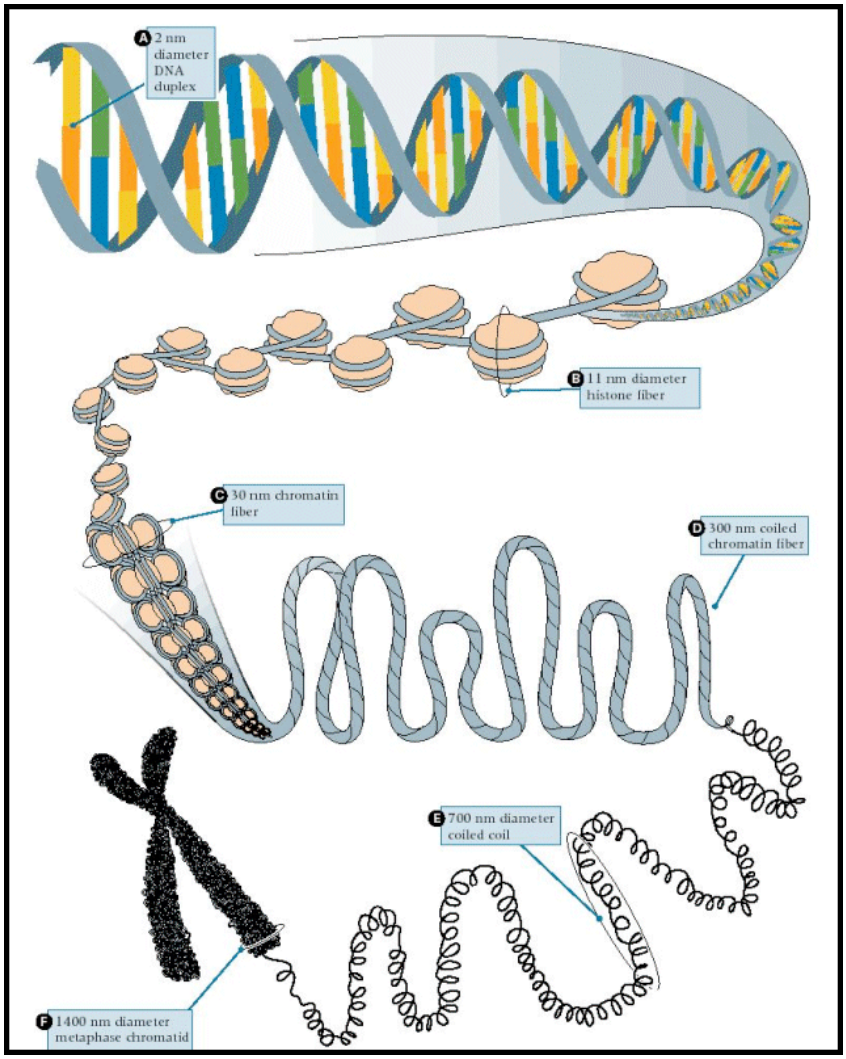
36. Chen, W., Tabor, S. & Struhl, K. Distinguishing between mechanisms of eukaryotic transcriptional activation with bacteriophage T7 RNA polymerase. *Cell* **50**, 1047–1055 (1987).
37. Zhu, Z. & Thiele, D.J. A specialized nucleosome modulates transcription factor access to a *C. glabrata* metal responsive promoter. *Cell* **87**, 459–470 (1996).
38. Zeevi, D. *et al.* Compensation for differences in gene copy number among yeast ribosomal proteins is encoded within their promoters. *Genome Res.* **21**, 2114–2128 (2011).
39. Field, Y. *et al.* Gene expression divergence in yeast is coupled to evolution of DNA-encoded nucleosome organization. *Nat. Genet.* **41**, 438–445 (2009).
40. Tirosh, I. & Barkai, N. Two strategies for gene regulation by promoter nucleosomes. *Genome Res.* **18**, 1084–1091 (2008).
41. Fan, X. *et al.* Nucleosome depletion in yeast terminator regions is not intrinsic and can occur by a transcriptional mechanism linked to 3′ end formation. *Proc. Natl. Acad. Sci. USA* **107**, 17945–17950 (2010).
42. Lantermann, A.B. *et al. Schizosaccharomyces pombe* gene-wide nucleosome mapping reveals positioning mechanisms distinct from those of *Saccharomyces cerevisiae*. *Nat. Struct. Mol. Biol.* **17**, 251–257 (2010).
43. Tsankov, A.M., Thompson, D.A., Socha, A., Regev, A. & Rando, O.J. The role of nucleosome positioning in the evolution of gene regulation. *PLoS Biol.* **8**, e1000414 (2010).
44. Tsankov, A., Yanagisawa, Y., Rhind, N., Regev, A. & Rando, O.J. Evolutionary divergence of intrinsic and trans-regulated nucleosome positioning sequences reveals plastic rules for chromatin organization. *Genome Res.* **21**, 1851–1862 (2011).
45. Zhang, Z. *et al.* A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome. *Science* **332**, 977–980 (2011).
46. Wippo, C.J. *et al.* The RSC chromatin remodelling enzyme has a unique role in directing the accurate positioning of nucleosomes. *EMBO J.* **30**, 1277–1288 (2011).
47. Badis, G. *et al.* A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Mol. Cell* **32**, 878–887 (2008).
48. Floer, M. *et al.* A RSC/nucleosome complex determines chromatin architecture and facilitates activator binding. *Cell* **141**, 407–418 (2010).
49. Rippe, K. *et al.* DNA sequence- and conformation-directed positioning of nucleosomes by chromatin-remodeling complexes. *Proc. Natl. Acad. Sci. USA* **104**, 15635–15640 (2007).
50. van Vugt, J.J. *et al.* Multiple aspects of ATP-dependent nucleosome translocation by RSC and Mi-2 are directed by the underlying DNA sequence. *PLoS ONE* **4**, e6345 (2009).
51. Whitehouse, I. & Tsukiyama, T. Antagonistic forces that position nucleosomes *in vivo*. *Nat. Struct. Mol. Biol.* **13**, 633–640 (2006).
52. Whitehouse, I., Rando, O.J., Delrow, J. & Tsukiyama, T. Chromatin remodelling at promoters suppresses antisense transcription. *Nature* **450**, 1031–1035 (2007).
53. Hartley, P.D. & Madhani, H.D. Mechanisms that specify promoter nucleosome location and identity. *Cell* **137**, 445–458 (2009).
54. Gkikopoulos, T. *et al.* A role for Snf2-related nucleosome-spacing enzymes in genome-wide nucleosome organization. *Science* **333**, 1758–1760 (2011).
55. Hennig, B.P., Bendrin, K., Zhou, Y. & Fischer, T. Chd1 chromatin remodelers maintain nucleosome organization and repress cryptic transcription. *EMBO Rep.* **13**, 997–1003 (2012).
56. Pointner, J. *et al.* CHD1 remodelers regulate nucleosome spacing *in vitro* and align nucleosomal arrays over gene coding regions in *S. pombe*. *EMBO J.* **31**, 4388–4403 (2012).
57. Shim, Y.S. *et al.* Hrp3 controls nucleosome positioning to suppress non-coding transcription in eu- and heterochromatin. *EMBO J.* **31**, 4375–4387 (2012).
58. Celona, B. *et al.* Substantial histone reduction modulates genomewide nucleosomal occupancy and global transcriptional output. *PLoS Biol.* **9**, e1001086 (2011).
59. Gossett, A.J. & Lieb, J.D. *In vivo* effects of histone H3 depletion on nucleosome occupancy and position in *Saccharomyces cerevisiae*. *PLoS Genet.* **8**, e1002771 (2012).
60. Ito, T., Bulger, M., Pazin, M.J., Kobayashi, R. & Kadonaga, J.T. ACF, an ISW1-containing and ATP-utilizing chromatin assembly and remodeling factor. *Cell* **90**, 145–155 (1997).
61. Drew, H.R. Reconstitution of short-spaced chromatin from the histone octamer and either HMG14,17 or histone H1. *J. Mol. Biol.* **230**, 824–836 (1993).
62. Fan, Y. *et al.* Histone H1 depletion in mammals alters global chromatin structure but causes specific changes in gene regulation. *Cell* **123**, 1199–1212 (2005).
63. Hashimoto, H. *et al.* Histone H1 null vertebrate cells exhibit altered nucleosome architecture. *Nucleic Acids Res.* **38**, 3533–3545 (2010).
64. Oberg, C., Izzo, A., Schneider, R., Wrange, O. & Belikov, S. Linker histone subtypes differ in their effect on nucleosome spacing *in vivo*. *J. Mol. Biol.* **419**, 183–197 (2012).
65. Rhee, H.S. & Pugh, B.F. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature* **483**, 295–301 (2012).
66. Adelman, K. & Lis, J.T. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat. Rev. Genet.* **13**, 720–731 (2012).
67. Chang, G.S. *et al.* Unusual combinatorial involvement of poly-A/T tracts in organizing genes and chromatin in *Dictyostelium*. *Genome Res.* **22**, 1098–1106 (2012).
68. Vaillant, C. *et al.* A novel strategy of transcription regulation by intragenic nucleosome ordering. *Genome Res.* **20**, 59–67 (2010).
69. Yen, K., Vinayachandran, V., Batta, K., Koerber, R.T. & Pugh, B.F. Genome-wide nucleosome specificity and directionality of chromatin remodelers. *Cell* **149**, 1461–1473 (2012).

1 A General Introduction to Nucleosomes and Nucleosome Positioning

1.1 Nucleosomes: the Building Blocks of Chromatin

Chromatin is the complex of DNA and cellular proteins which form eukaryotic chromosomes. It is composed of an elementary repeating unit called the nucleosome, which is the major factor of DNA packaging in eukaryotic genomes (Figure 1.1).

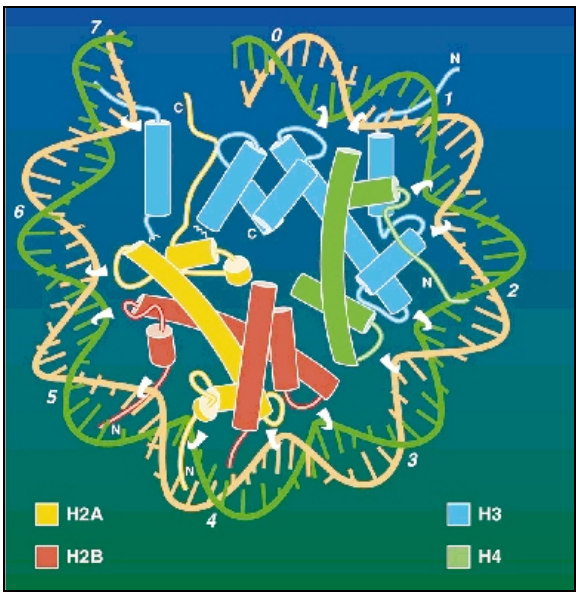
Figure 1.1: A hierarchical view of chromatin structure. Reproduced figure (Hartl & Jones, 1998).



Nucleosomes are DNA-protein complexes, which are comprised of a *core particle* of 1.6 left-handed turns of DNA (roughly 146 bp) wound around a protein complex called the *histone octamer* (Figure 1.1(B)). The histone octamer is a set of 8 basic proteins, which are among the most well conserved proteins known in eukaryotes. It is comprised of a central tetramer, $(H3/H4)_2$, flanked by two H2A/H2B

dimers (Figure 1.2). The structure of a single histone molecule includes three major α helices with positively-charged loops protruding at the N-terminals.

Figure 1.2: Top-level view of a nucleosome. Cylinders indicate alpha-helices; white hooks represent arginine/lysine tails. Reproduced figure (Rhodes, 1997)).



The DNA wrapped around the histone octamer is called the *core DNA* and the DNA joining adjacent nucleosomes is called *linker DNA*. Unlike core DNA, linker DNA exhibits great variability in length: anywhere between 8 to 200 bp. This variation in the length of linker DNA may be important for the diversity of gene regulation; however, chromatin structure formation is independent of the length of linker DNA (Kornberg & Lorch, 1999).

The constraint of the nucleosome on the DNA path forms the first level of higher-order packing, compacting DNA by a factor of ~6 (Lewin, 2000). An extra histone H1 (also called the linker histone) may also be present, clamping the DNA at the position at which it enters and leaves the histone core (Karrer & VanNuland, 1999; Satchwell & Travers, 1989; Widlund *et al.*, 2000).

The series of nucleosomes along a DNA sequence then coil into a helical array forming a fibre of ~30 nm (Figure 1.1(C)); this results in further compaction by a factor of ~40. In the recent crystal structure of the nucleosome (Luger *et al.*, 1997),

it had been reported that the basic tail of H4 protrudes extensively and makes contacts with acidic patches of H2A and H2B on neighbouring octamers; this implies a role for H4 in stabilizing higher level structures. Histone H1 is thought to appear mainly towards the middle of the 30 nm fibre where it may play a role in stabilizing chromatin interactions (Staynov, 2000). Specialised nucleosomes are also known, for example the centromere-specific nucleosomes, which contain a variant of histone H3 called CENP-A; these occur in a range of organisms from yeast to human (Smith, 2002). Many non-histone chromatin proteins also interact with histones to enable formation of higher-order structures. The fibre itself undergoes further levels of packaging resulting in compaction by a factor of ~1000 in interphase euchromatin and ~10,000 in heterochromatin (Figure 1.1(D-F)).

The structure of chromatin is dynamic. It exists in a number of distinct functional states which can often be characterised by the level of transcriptional activity. The dynamic transitions between these states occur through a range of post-translational modifications of the histone tails which includes acetylation and phosphorylation (Jenuwein & Allis, 2001). This forms the basis of the "histone code hypothesis" which states that the combinatorial nature of these modifications results in the generation of altered chromatin structures that mediate specific biological responses (Turner, 2000).

1.2 DNA-Protein Interactions in the Nucleosome Core Particle

The earliest concepts for the association of DNA and histones in the core particle came from image reconstruction analysis using electron micrographs (Klug *et al.*, 1980). At 20 Å resolution, a left handed helical ramp was apparent on the octamer surface and proposals were made for how the DNA-protein interactions might occur. Since then, X-ray crystallography has helped to advance understanding of the DNA-protein interactions involved in the nucleosome core particle. Milestones included the solving of the nucleosome structure at 7 Å resolution (Uberbacher & Bunick, 1985), which reconfirmed the initially inferred arrangement of histones and DNA. This led to the highest resolution structures of the nucleosome core particle to date at 2.8 Å (Luger *et al.*, 1997) and 1.9 Å (Davey *et al.*, 2002).

The high-resolution structure of the core-particle firstly revealed that the core particle had a pseudo-dyad[1] axis of symmetry: 1 bp sat on the dyad axis of the octamer. It further revealed in fine detail that the histone-DNA interactions were confined towards the phosphodiester backbone of the DNA strand (Luger *et al.*, 1997). Arginine/lysine-rich tails, protruding from the core histones, made "hook-like" contacts every 10 bp where the minor groove of the double-helix faced inwards. The histone-DNA contacts were non-base-specific and included predominantly salt-bridges and H-bonds as well as non-polar contacts with DNA sugars.

The 10 periodic contact feature of the DNA backbone was suggested much earlier. It was suggested, for example, when 10 bp-phased digestion patterns were observed upon using the enzyme DNase I[2] to cut nucleosome-bound DNA (Wang,

[1] The central axis of the histone octamer is herein referred to as the dyad axis.
[2] DNase I is an endonuclease, which breaks phosphodiester bonds within DNA.

1982). The observed cutting periodicity of 10 bp, which is "in phase" with the helical periodicity of DNA, forms the basis of many computational approaches aimed at finding nucleosome rotational positioning signals (Section 1.9).

The helical periodicity of DNA is not constant as it traverses around the histone octamer. For example, experiments using hydroxyl-radical cleavage of nucleosome-bound DNA showed that the helical periodicity was 10.0 bp/turn in the vicinity of the dyad axis and 10.7 bp/turn towards the ends of the nucleosome (Puhl & Behe, 1993). Most experimental evidence for B-DNA in solution suggests that it has a helical periodicity of 10.5–10.6 bp in solution (Wolffe, 1998). This variation in DNA periodicity along the core particle is thought to be a consequence of local histone-DNA interactions.

1.3 The Concept of Nucleosome Positioning

Nucleosome positioning has been proposed to be a potential mechanism for regulating gene expression, providing the view that nucleosomes could play important roles in addition to organizing higher order chromatin structures in eukaryotic cells. The term 'positioning' refers to a pre-determined organization of nucleosomes on a DNA sequence. In contrast, in a random arrangement of nucleosomes, all DNA sequences will have an equal probability of binding histones (Sinden, 1994). This gives rise to the idea that the local DNA structure, which is affected by the underlying DNA sequence, may play a role in positioning nucleosomes.

Two kinds of DNA structural patterns may thus be envisioned to direct nucleosome positioning: those that strongly favour nucleosome formation and those that strongly obstruct it. Nucleosome positioning can help to either selectively expose functionally important DNA sequences by constraining their locations to the linker region or impede accessibility to functionally important sequences by constraining their location to within the core particle. This can impose another level of regulation in gene expression, for instance, by controlling the accessibility of binding sites available to RNA polymerases or specific transcription factors. Two kinds of DNA structure-based nucleosome positioning have been described previously and these will be discussed next (Sections 1.4, 1.5).

1.4 An Introduction to Nucleosome Rotational Positioning

Rotational positioning determines which side of a DNA double helix surface will face and contact the histone octamer; this kind of positioning has been attributed to intrinsically curved DNA. The theory that a nucleosome will fit an intrinsically curved DNA is that the DNA is already in a preferred physical conformation to allow it to easily wrap around the octamer surface.

This section will firstly introduce the physical basis of DNA which results in intrinsic curvature and then describe how this relates to rotational positioning preferences for nucleosomes.

1.4.1 Intrinsic DNA curvature: Bending based on 10-phased [A] tracts

Intrinsically curved DNA is thought to be a consequence of permanent bends in a DNA sequence. This was first proposed when it was noticed that a 414 bp piece of kinetoplast DNA from *Crithidia fasciculata* displayed limited or retarded migration compared to other sequence fragments of equal length in acrylamide gel but migrated normally in agarose gel (Marini *et al.*, 1983). This anomalous migration was attributed to the size of the pores in the respective gels: in acrylamide gels, pore sizes vary between 1-8 nm whereas pore sizes in agarose gels vary between 40-400 nm. It was proposed that a permanent bend or curvature in the kinetoplast sequence was probably what caused the fragment to get stuck in the smaller size pores of the acrylamide gels.

The sequence motif that caused the permanent bends was mapped using the circular permutation assay (Wu & Crothers, 1984). In this procedure, various 241 bp-long restriction fragments, of the 414 bp-long kinetoplast DNA, were prepared and

cloned as dimers. The length of 241 bp was chosen as this is greater than the persistence length of DNA[3]. The dimerized fragments were then run on an acrylamide gel and scanned for the fragment causing the shortest end-to-end migration distance (this region contained the permanent bend). This experiment concluded the retarded migration property to be an effect of 10 bp-phased runs of $CA_{4-5}T$ in the kinetoplast DNA. This work led Wu *et al* to propose the *junction model* for DNA bending; this predicts that the poly(dA)·poly(dT) tracts, within the 10 bp-phased $CA_{4-5}T$ motifs, adopt a non-B-DNA helix called heteronomous DNA (Arnott *et al.*, 1983). It proposes that permanent bends are located at the junction between this kind of DNA and regular B-DNA.

An alternative model to explain how phased-A tracts caused permanent bending was proposed later called the *wedge model* (Ulanovsky *et al.*, 1986). In this assessment, "bend angles" were calculated by measuring the efficiency of ligation of small DNA fragments into closed circles. This model predicts that the bends are not located at the junction between 2 kinds of DNA structure but within the [AA] dinucleotides themselves.

Parameters estimated from X-ray analysis of DNA structure have also been used to explain how phased-A tracts could cause intrinsic DNA curvature. From X-ray crystal structures, 2 variables are considered important for the relative motion of DNA base pairs: *roll* and *slide* (Calladine & Drew, 1992). *Roll* describes the opening of base pairs towards the major or minor groove of the double helix. A positive roll value indicates a tendency to open up towards the minor groove whereas a negative roll value indicates a tendency to open up towards the major groove in the opposite direction; typical values for DNA bases range between $+20^0$ to -10^0. *Slide* refers to

[3] The persistence length of DNA is 150 bp, the minimum length at which random DNA is essentially linear: it cannot circularize.

the translation along the axis of the base pairs. Slide values, which are restricted by the sugar-phosphate chain, range from +2 Å to -1 Å. Estimates of roll angle from X-ray structure analysis predict [AA/TT], [AT] and [GA/TC] dinucleotides to be stable at low roll (0^0) and low slide (0 Å) (El Hassan MA & Calladine, 1997) making their overall conformation very restricted. On the other hand, dinucleotides such as [GC/GC], [CG/CG] and [GG] dinucleotides are predicted to exhibit a wide range of roll angles (-10^0 to 20^0) making their conformation "bistable" or "context-dependent". For the phased A-tract bending, this suggests that the [AA] dinucleotides prefer to align their side of the minor groove towards the centre of curvature because of their restricted low roll configuration and the [GC] dinucleotides prefer to align the major groove away from the centre of curvature because of their bistable configuration (more on this below; Section 1.4.2).

The latest evidence that tries to explain how phased-A tracts result in bending comes from NMR studies (MacDonald *et al.*, 2001). This estimates a total of 19^0 bending in phased A-tracts. Of this, 4^0 occurs at the 5'end of the A-tract, 5^0 occurs within the A-tract itself and 10^0 occurs at the 3' end of the A-tract.

1.4.2 Intrinsic DNA curvature and the initial assessment of nucleosome rotational positioning

A rotational preference for a circular piece of DNA sequence has been described as a bias towards aligning a specific face of the DNA surface towards the direction of curvature and aligning a specific face away from the direction of the curvature (Drew & Travers, 1985). To study the rotational preferences of 10 bp-phased [A] tract sequences, a 169 bp sequence, containing phased [A]-tracts, was circularly ligated and

digested with DNase I[4] (Drew & Travers, 1985). The [GC]-tracts were seen to be easily digested by DNase I and therefore more likely to face away from the circle. On the other hand, the phased [A]-tracts were more likely to be oriented towards the circle and thus protected from DNase I digestion. This observation was consistent with the X-ray crystal structure explanation of [A]-tract DNA bending discussed in the previous section (Section 1.4.1). As part of the same experiment, the same sequence was reconstituted onto a histone octamer *in vitro.* Digesting this reconstituted nucleosome with DNase I showed the same rotational preferences as for the circularized DNA: the phased [A]-tracts of the sequence were seen to face in towards the histone octamer. A later study addressed the optimal number of [A] nucleotides required for [A]-tract bending (Koo *et al.*, 1986). The approach used gel anomaly analysis of several lengths of [A] nucleotides in 10 bp-phased [A]-tract sequences. This study showed that 3–5 [A] nucleotides, phased at 10 bp, resulted in optimal curvature.

Further analysis of rotational positioning of DNA sequences on histone octamers was carried out by cloning a library of 177 nucleosome core particle sequences from chicken genomic DNA and subsequently analysing its dinucleotide periodicity (this dataset is discussed again subsequently in Section 1.8.1) (Satchwell *et al.*, 1986). The sequence lengths in the final dataset, however, were not constant, most probably due to biases in micrococcal nuclease (MNase[5]) cutting specificity (Section 1.8.1). The lengths ranged from 142 to 149 bp with an average length of 145 (±1.5) bp. To deal with this uncertainty, the analysis was carried out using 3 bp-averaged representations of the data. Also, the authors had to shift all sequences,

[4] DNase I interacts with the surface of the minor groove and bends the DNA molecule away from the enzyme.
[5] Micrococcal nuclease is both an endonuclease and an exonuclease, which can break the phosphodiester bonds in linker DNA and remove nucleotides from the ends of the DNA molecule respectively.

which were not of length 145 bp, a few base pairs until a central reference point of 73.25 was obtained. Fourier analysis of the dinucleotides in the dataset showed 10 periodic patterns of [AA/TT] and [GC]. These 2 motifs were furthermore seen to occur phased at 5 bp from each other, reminiscent of the A-tract bent sequences discussed in the previous section.

In the same study (Satchwell *et al.*, 1986), the 3 bp-averaged positions of dinucleotide motifs were compared with the co-ordinates of the DNA sequence which faced the octamer in the nucleosome X-ray crystal structure available at that time (Richmond *et al.*, 1984). This showed a pattern for phased A-tracts to face the octamer a few turns symmetrically away from the dyad axis of the nucleosome core particle but not at the dyad itself. In the X-ray crystal structure of the nucleosome, the minor groove also faced away from the dyad axis (Section 1.2). This result also agrees with the previous discussion that there are 2 kinds of DNA helical periodicities at the dyad and end positions respectively (Section 1.2).

1.4.3 Further evidence to support nucleosome rotational positioning

Since the initial assessment of nucleosome rotational positioning, a big trend was to chemically synthesise DNA sequences with optimised rotational preferences for forming reconstituted nucleosomes *in vitro*. For example, sequences having repeats of the motif [TATAAACGCC] were shown to ligate more efficiently into a circle compared to random DNA (Widlund *et al.*, 1999). This sequence was shown to bind nucleosome core particles *in vitro* ~350 fold higher than random DNA. A few naturally phased A-tract sequences are also known to favour nucleosome reconstitution *in vitro*, for example the 5S RNA gene of *Xenopus laevis* (Tomaszewski & Jerzmanowski, 1997).

Analysis of whole genomic sequences has also shown that they may contain enriched phased A-tract bending motifs for positioning nucleosomes. For example, Fourier analysis of *Caenorhabditis elegans* and *Saccharomyces cerevisiae* showed enrichment of [AA] motifs at 10.2 bp (Widom, 1996); the same pattern was not seen in a prokaryotic genome. A different approach to analyzing whole genomic sequences is the SELEX protocol (Widlund *et al.*, 1997). This procedure works by starting off with a random pool of genomic sequences and performing a number of rounds of PCR, each time amplifying sequences based on their affinity to bind histones. This approach found [A]-tract bending sequences in *Methanothermus fervidus*, which belongs to a branch of the archaeal kingdom that contains histone like proteins (*Euryarchaeota*) (Bailey *et al.*, 2000). The same patterns were not found in *Crenarchaeota*, a branch of the archaeal kingdom which does not contain histones. This led to the suggestion that the evolution of eukaryotic genome sequences most likely originated in the archaea, before the split of the eukaryotic lineage.

1.4.4 Nucleosome rotational positioning and DNA regulatory regions

Generally, chromatin structure provides a repressive environment for transcription. The evidence for this comes from observations of increased transcription levels of prokaryotic RNA polymerases in histone-depleted eukaryotic cells compared to their levels in normal eukaryotic cells (Gonzalez & Palacian, 1989). Prokaryotic RNA polymerases have traditionally been used in such analyses since they do not require specific transcription factors as do eukaryotic RNA polymerases (Wolffe, 1998). One of the ways eukaryotic cells are understood to overcome nucleosome barriers to permit transcription is through the activity of ATPase-based remodelling complexes (Wolffe & Guschin, 2000). An example is the SWI/SNF complex, which is thought

to disrupt the rotational positioning of nucleosomes as suggested from the loss of 10 bp-phased DNase I cleavage patterns (Lorch *et al.*, 1998).

The indication for nucleosome rotational positioning provided an incentive to map naturally bent DNA near important genomic sequences and assess whether these bends could position nucleosomes (Bash *et al.*, 2001; Nair, 1998; Pruss *et al.*, 1994; Wada-Kiyama & Kiyama, 1996; Wada-Kiyama *et al.*, 1999).

For example, the circular permutation assay (Section 1.4.1) was used to map bend sites in the 3,000 bp promoter region of the human oestrogen receptor gene (Wada-Kiyama *et al.*, 1999). A total of 5 bend sites were found using the circular permutation assay; [A]-tract bending was observed for 3 of these sites. Nucleosome positioning at one of these bend sites was then analysed in detail. These were mapped by firstly digesting the clone with MNase to isolate core particles followed by digestion with 2 different restriction enzymes, whose restriction sites were known on the clone. This showed that the position of the bend appeared 10–30 bp away from the experimentally-predicted location of the nucleosome dyad axis. Therefore, it seemed likely that the specific bent site could help to direct nucleosome positioning. Nucleosome mapping to an intrinsically bent site was shown previously as well in the human β globin locus (Wada-Kiyama & Kiyama, 1996).

A few specific cases are known where positioned nucleosomes are important for protein signal recognition. An example of this is the hormone responsive element (HRE) of the mouse mammary tumour virus (MMTV) promoter (Pina *et al.*, 1990). Footprinting[6] analysis showed that the sequence of HRE was able to precisely position nucleosomes both *in vivo* and in reconstituted chromatin. It was then shown that nuclear factor 1 (*NFI*), one of the transcription factors for this promoter, was not

[6] This technique identifies the site of protein-binding on DNA by determining which phosphodiester bonds are protected from cleavage by DNase I

able to bind to the promoter when it was wrapped in a nucleosome. Hormone receptor binding to the MMTV nucleosome was seen to shift the rotational position of the nucleosome rather than causing it to dissociate completely; this was detected as greater accessibility of the promoter-proximal end to exonuclease III digestion. Thus, hormone receptor binding could act as a primary switch by shifting the rotational setting of the nucleosome to permit *NF1* binding. Another example is the binding site of the human immunodeficiency virus (HIV)-encoded integrase enzyme: DNA distortion studies have shown that this enzyme recognises specific bends within a nucleosome core particle (Pruss *et al.*, 1994).

1.5 An Introduction to Nucleosome Translational Positioning

Translational positioning determines where a histone octamer will be positioned along a long stretch of DNA; "long", in this case, refers to a length longer than the core particle length (~146 bp). The theory behind this kind of positioning is that certain regions of a long DNA sequence may be much worse or much better than random DNA in their ability to wrap a histone octamer. Two kinds of DNA structural features may be important in determining the translational position of a nucleosome:

- Highly rigid DNA – DNA, whose structural conformation is very restricted, compared to random DNA, will be more difficult to bend around a histone octamer. Therefore, such kind of DNA can be expected to repel nucleosome formation.
- Highly flexible DNA - The conformation of highly flexible DNA is such that it offers least resistance to being bent and wrapped around a histone octamer. Thus, DNA, which is significantly more flexible than random DNA sequences, may position nucleosomes more readily. Flexible DNA is different to bent DNA previously described (Section 1.4.1) in that it offers low resistance to being wrapped around a histone octamer whereas bent DNA is a permanent feature of the DNA molecule.

1.5.1 DNA sequences that repel nucleosome formation

Sequences that resist nucleosome formation may do so because they tend to form some other kind of DNA secondary structure unfavourable for wrapping around a nucleosome. They might also contain signals to bind a different cellular protein, which would compete with the histone octamer for the same position. Initial

nucleosome reconstitution experiments, using salt dialysis, had reported a lack of success in reconstituting nucleosomes using poly(dA)·poly(dT) / poly(dG).poly(dC) sequences (Rhodes, 1979; Simpson & Shindo, 1979). Although it was not clear why such sequences would disfavour nucleosome formation, Rhodes *et al* suggested that the high salt conditions used in the reconstitution procedure could have caused the poly(dA)·poly(dT) sequences to form heteronomous DNA, a triple-strand DNA structure (Arnott *et al.*, 1983). Poly(dG).poly(dC) sequences were also known to easily adopt A-DNA conformation (Arnott & Selsing, 1974) so this could have been a possibility for their inability to reconstitute into nucleosomes using the high-salt experimental conditions.

In another nucleosome reconstitution experiment, it was also observed that tracts of poly(dA)·poly(dT) and poly(dG).poly(dC) were not present towards the dyad axis (Drew & Travers, 1985). However, poly(dA)·poly(dT) tracts appeared towards the ends of the core DNA sequences suggesting that they may have an influence on the translational setting of the histone octamer (Satchwell *et al.*, 1986). The basis for translational positioning was not clear at this point; a recent study, however, examined the translational and rotational positioning properties of a simple 20 bp-repeating sequence (Negri *et al.*, 2001). The approach was to study the effects of subtle changes to the original sequence by mapping the changes to rotational and translational positions using hydroxyl-radical and exonuclease mapping respectively. The main conclusion was that the sequence distortions which affected the rotational preferences of the core particle were not the same ones which affected the translational position. The exact features which resulted in translational positioning, however, were not confirmed but it was suggested that the exact sequence contexts of [GA] and [CT] dinucleotides could be important.

Why long runs of poly(dA)·poly(dT) might repel nucleosome formation is still unclear. However, one explanation, using X-ray crystal analysis, predicts A·T base pairs to have high propeller twist[7] (Nelson *et al.*, 1987). This would result in maximal base-stacking (the interaction of adjacent base pairs) in poly(dA)·poly(dT) sequences resulting in an overall rigid stretch of DNA. [AA/TT] dinucleotides were also discussed earlier to show restricted conformation in X-ray crystallography studies (Section 1.4.1). This may make it difficult to bend poly(dA)•poly(dT) sequences to easily fit around a histone octamer.

Expansion of [CCG] repeats, which are known to cause fragile X syndrome, have also been studied in relation to nucleosome positioning (Wang *et al.*, 1996). Using competitive nucleosome reconstitution and electron microscopy, it was shown that >50 repeats of [CCG] blocks tended to exclude nucleosome formation. Such sites, visible in patients suffering from fragile X syndrome, were referred to as "fragile" loci as they stained poorly and were hotspots for DNA strand breakage. It was possible that [CCG] repeats formed some other kind of secondary structure: the lack of nucleosomes could account for the high frequency of DNA strand breaks. The exact mechanism for extensive CCG repeats in excluding nucleosome formation is still unclear.

Cao *et al* had performed a negative-SELEX experiment on mouse genomic DNA to yield an enriched quantity of sequences that repel nucleosome formation (Cao *et al.*, 1998). 35% of the sequences finally isolated had long repeats of [TGGA] and the affinity of these were half that of background DNA.

[7] Propeller twist is a property of a single base pair which describes the angle between the plane of the paired bases.

1.5.2 DNA sequences that favour nucleosome formation

Expanded blocks of [CTG] have been shown to be strong positioning signals for binding nucleosomes (Wang & Griffith, 1995). This motif had been previously shown to form expanded blocks downstream of the myotonic dystrophy gene in affected patients (Mahadevan *et al.*, 1992). Such regions were seen to bind a large number of nucleosomes using electron microscopy. An *in vitro* nucleosome reconstitution experiment showed that 2 DNA sequences, having 75 and 130 [CTG] repeats respectively, formed nucleosomes 6 and 9 times more strongly compared to the 5S RNA gene (a naturally occurring nucleosome-positioning sequence containing 10 bp-phased [A]-tracts) (Wang & Griffith, 1995). A study involving DNase I digestion of trinucleotides has also shown [CTG] trinucleotides to have one of the highest cutting rates and therefore to be amongst the most flexible trinucleotides (Brukner *et al.*, 1995). So according to the DNase I digestion results, the high flexibility of [CTG]-expanded regions may lead to a relatively "easy" fit for binding nucleosomes. However, according to the analysis of the chicken nucleosome data, [CTG] motifs did not show any kind of rotational positioning preferences, i.e. to face inwards or outwards in the structure of the core particle (Satchwell *et al.*, 1986). This suggests that [CTG] may show preferential nucleosome binding only when it is present in dense clumps: its overall density along a DNA sequence and not its rotational preference may influence its strong nucleosome-binding feature.

SELEX enrichment of core DNA in the mouse genome found some other possible nucleosome-positioning motifs, all of which could not be explained by phased [A]-tract motifs (Widlund *et al.*, 1997). This study found some cases of phased runs of 3-4 adenines ([A]-tract bending), multiple [CA] repeats, phased [TATA] tetranucleotides and one sequence having [CAG] repeats. However,

fluorescence *in situ* hybridization showed these sequences to strongly localise to centromeric DNA; some of the sequence motifs were also known centromeric satellite repeats. Such repeats may not represent the majority of nucleosome-binding sequences in the genome as centromeric nucleosomes contain specialised nucleosomes that have variant histones (Smith, 2002). Furthermore, a recent study showed that the exact histone variant in addition to the DNA sequence may be a factor in positioning nucleosomes (Bailey *et al.*, 2002).

1.5.3 Nucleosome translational positioning and DNA regulatory regions

As mentioned earlier, nucleosomes are considered a repressive environment for transcription (Section 1.4.4). To overcome this, eukaryotic cells also contain ATPase-based remodelling complexes which are understood to shift the translational positioning of nucleosomes, for example NURF complexes in Drosophila (Hamiche *et al.*, 1999; Kang *et al.*, 2002). These are thought to induce sliding of nucleosomes as they do not disrupt the 10 bp-phased DNase I digestion patterns.

Understanding of the role of translational nucleosome positioning in repressing transcription has come from the use of *in vitro* transcription systems (Wolffe, 1998). Such studies ask if transcription can still occur *in vitro* following nucleosome reconstitution. The general outcome is that if histone assembly takes place first, transcription activity is inhibited. Of course, this system is unlikely to represent what happens in eukaryotic cells *in vivo* as it is difficult to mimic the multitude of transcription factors, which are actively involved in the process. An experiment, using an *in vitro* transcription system, showed that Alu repeats positioned histones over and next to promoter elements, which are critical for its transcription

activity (Englander *et al.*, 1993). The poly [A] linker region of Alu sequences was proposed to exclude translational positioning by a histone octamer.

1.6 Regions of Phased Nucleosomes

One of the consequences of nucleosome positioning may be genomic segments having 'phased nucleosomes': in this case, a constant length of linker DNA is maintained throughout a specific segment of genomic sequence. Possible models for demarcating such segments have been proposed (Kiyama & Trifonov, 2002):

- A perfect positioning model – The positions for all nucleosomes are defined in a genomic segment.
- A partial positioning model – Certain positions in a genomic segment are designated for nucleosome formation. The alignment of other nucleosomes is influenced by the initial allocation of these key positions.

A crude method of detecting nucleosome phasing in a genomic clone is by digesting it with micrococcal nuclease and observing the digested products using gel electrophoresis. If the bands produced by electrophoresis produce a unique band, it suggests that the linker lengths are roughly equal and that a specific phase is maintained. Conversely, "out of phase" nucleosomes yield a number of bands of varying lengths. Nucleosome-phasing was observed in a few randomly selected chicken genomic DNA clones using this method (Liu & Stein, 1997). This study concluded that phased regions (<2k bp) alternated with randomly-positioned regions in the sampled clones; the phased regions were reported to show 210 bp-phased nucleosomes. Possible underlying sequence factors were proposed in one of the phased regions, which contained a gene. These included a run of 10 [A] residues in the linker DNA between 2 specific nucleosomes (possible translational positioning motif) and apparently 10 bp-phased [VWG] motifs (Section 1.9.3; a motif that could affect rotational positioning).

1.7 Strength of Nucleosome Positioning Sequences In Vivo

Two very important problems have been looked at previously concerning the strength of nucleosome positioning sequences *in vivo*. The first was to estimate what proportion of genome sequences might be constrained for packaging nucleosomes. The second problem was to answer how efficient these sequences were at binding octamers compared to artificial sequences.

The first question was answered using competitive nucleosome reconstitution in which a library of random natural genomic mouse DNA sequences and a library of chemically synthetic DNA (Lowary & Widom, 1997) were made to compete for binding limiting amounts of histone octamer. The conclusion was that only 5% of the total genomic library was enriched to bind histones with a free energy of reconstitution higher than the synthetic library.

To address the second problem about the strength of naturally occurring motifs, a set of the strongest possible motifs in the whole mouse genome was enriched and analysed using SELEX enrichment (Widlund *et al.*, 1997). The free energies of these sequences were compared with artificial sequences, which were similarly enriched for nucleosome-binding using SELEX enrichment (Thastrom *et al.*, 1999). The first and second strongest sequences in the entire mouse genome were seen to have 6 fold and 34 fold lower affinities respectively for binding octamers than the random pool of synthetic DNA. It was concluded that even the strongest binding natural sequences were not evolved to be the most energetically favourable possible.

1.8 Experimentally Mapped Nucleosome Datasets

Two databases of experimentally-mapped nucleosome sequences were available during the course of work described in this thesis. Sequences in both databases, however, suffer from experimental limitations which hinder the precise mapping of the dyad axis.

1.8.1 Database of chicken core DNA sequences

The database of chicken core DNA, which was introduced earlier (Section 1.4.2) (Satchwell *et al.*, 1986) (177 sequences), was kindly made available by Andrew Travers. To isolate core DNA, MNase digestion was performed on DNA extracted from chicken red blood cells. This was followed by a further deproteination step to remove H5 (the chicken equivalent of the linker histone H1 in human). This resulted in 239 sequences, which were cloned using an M13 vector, and sequenced. However, many of the cloned sequences were finally discarded: these included those that were less than 142 bp and those that contained a double-length insert of roughly 290 bp. The sequence lengths in the final database ranged from 142 to 149 bp with an average length of 145 (±1.5) bp.

The length differences could be partly attributed to the cutting specificities of MNase. It prefers cutting pA and pT faster than pC or pG (Bellard *et al.*, 1989) resulting in an accuracy of ±3 bp in determining the translational positioning of the core particle (Hager & Fragoso, 1999). However, the authors reported that the A+T content in the core particles were the same as those in bulk chicken DNA (Satchwell *et al.*, 1986). Only a drop of 13% in TpA between core particle DNA and bulk chicken DNA was noticed that could be biased by MNase cutting specificity.

The authors also mention that this dataset did not necessarily represent the bulk of nucleosome positioning *in vivo* as one step of the isolation protocol, which involved removal of H1, "*allowed the exchange of histone octamers between DNA molecules*" (Satchwell *et al.*, 1986).

10 bp-phased [AA/TT] periodicity, along with 5 bp phase-shifted [GC], had been reported for this dataset (Section 1.4.2). Simple counting of [AA/TT] dinucleotide spacing (Figure 1.5, page 1-31) and multiple alignments of these sequences (Appendix A) were not sufficient to reproduce this result. The multiple sequence alignment in Appendix A, which is also sorted by pair wise identity, showed that the sequences were not highly similar to each other. A separate BLAST analysis (Altschul *et al.*, 1990)was also performed where each of the core DNA sequences was used to search for homologous members in the dataset (an "all against all" test; data not shown). This showed that these sequences were not highly similar to each other. This suggested that the reported periodicity was probably quite weak.

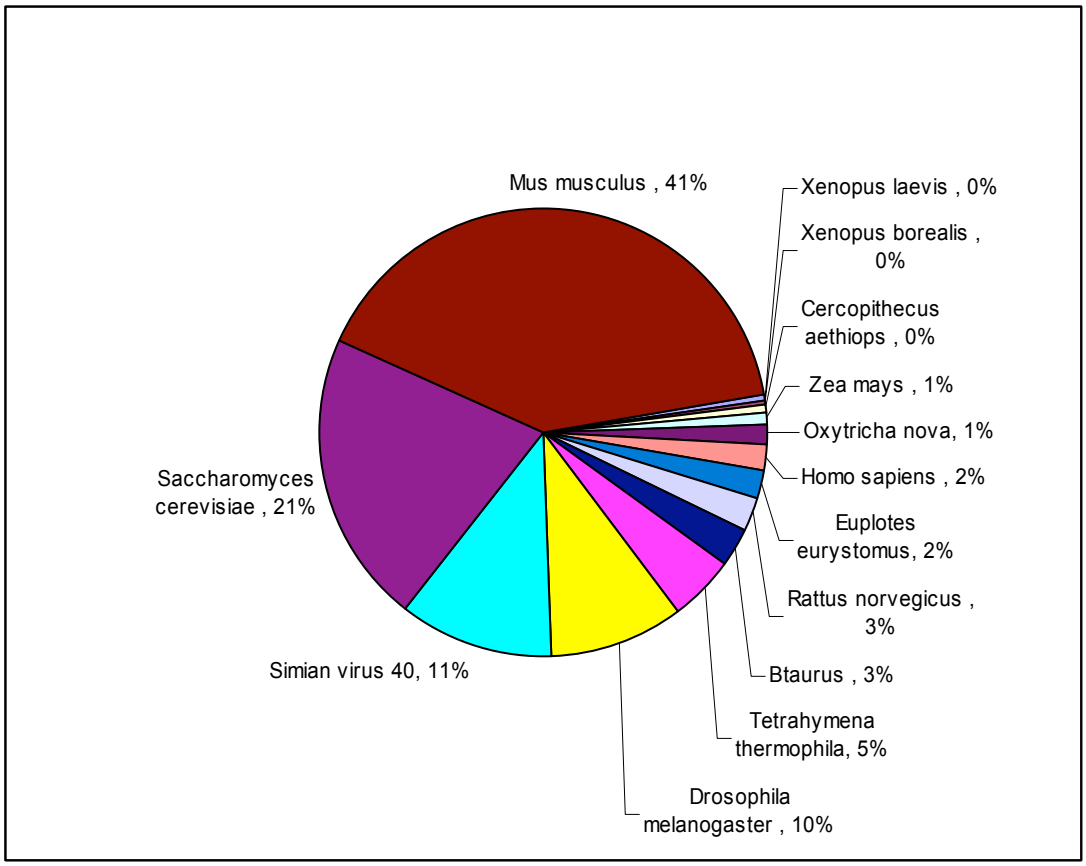
For some of the experiments performed in this thesis (Chapter 3 and Chapter 5), additional chicken genomic sequences were required which could be used as a background test set to these chicken core DNA sequences. Two chicken genomic clones were available for this purpose: AC092403 (144,369 bp) and AC120196 (202,027 bp).

1.8.2 Nucleosome database from mapping studies on various species

A second database of nucleosome sequences, which was publicly available (Levitsky *et al.*, 1999), essentially represented the same sequences from an earlier collection (Ioshikhes & Trifonov, 1993) and a more recent database of mouse nucleosomal sequences obtained using SELEX enrichment (Widlund *et al.*, 1997). A total of 193

sequences was present with the majority of sequences representing mouse and yeast data (Figure 1.3).

Figure 1.3: Organism sources of Levitsky et al's nucleosome sequence dataset (Levitsky *et al.*, 1999).



However, the length distribution of sequences was much more varied in this dataset compared to the mapped chicken sequences (Figure 1.4). The observed length variation necessarily resulted from the uncertainty of the technique used for nucleosome mapping. There were six main methods used, whose mapping accuracies are shown in Table 1.1 (Ioshikhes & Trifonov, 1993). The only technique unlisted in Table 1.1 is the SELEX protocol used to isolate many of the mouse nucleosome sequences: the lengths of these sequences ranged from 109 to 151 bp (average: 129 bp, standard deviation: 9 bp).

Figure 1.4: Length distribution of sequences in Levitsky *et al*'s nucleosome database.

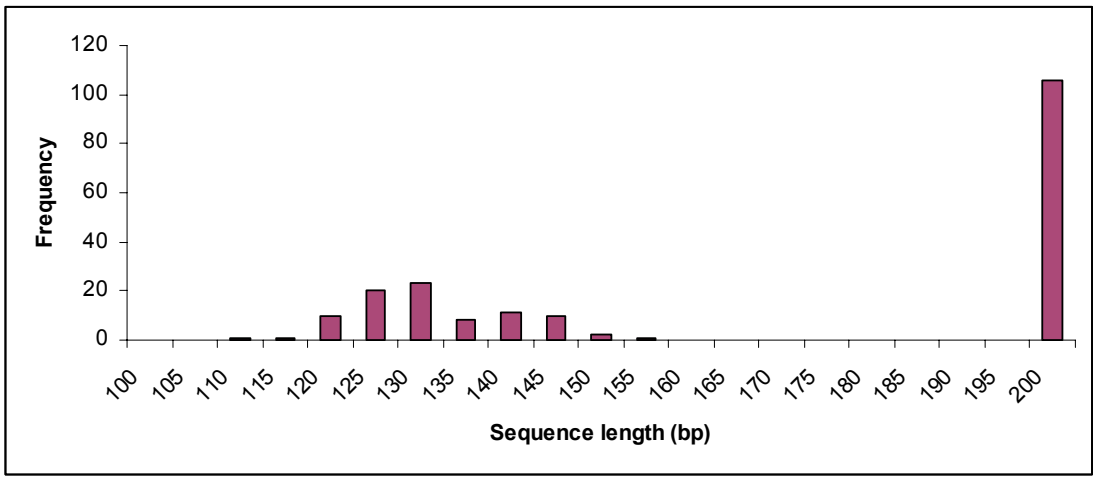


Table 1.1: Accuracy of different nucleosome mapping methods (Ioshikhes & Trifonov, 1993).

METHOD	MAPPING ACCURACY (bp)
MNase digestion of chromatin	>19
DNase I digestion of chromatin or reconstituted nucleosomes	10
Hydroxyl radical mapping	5
MNase digestion in combination with the cloning and sequencing of nucleosomal DNA sequences	5
DNase I digestion in combination with the highest possible accuracy	1
Exonuclease III with nuclease S1 digestion	1

The pair wise multiple sequence alignment of these sequences (Appendix A) showed that many of the mouse sequences were highly similar to each other (sequences 1-36 in the alignment). An "all against all" BLAST analysis also showed that these mouse sequences were highly similar to each other. However, they were more similar to the other sequences within the dataset compared to the chicken core DNA dataset (data not shown). The largely redundant mouse sequences were removed for any further analysis performed in this thesis. Unlike the chicken core DNA sequences, the sequence alignment of this dataset showed what appeared to represent phased [A]-tract motifs; these were in the first half of these sequences (Appendix A). [A]-tract bending was, therefore, more indicative in this dataset than

in the chicken nucleosome dataset (this is discussed again subsequently; Section 1.9.2).

1.9 Computational Approaches to Understanding Nucleosome Positioning in Other Laboratories

This section will briefly introduce some of the computational approaches that have been developed till now to predict nucleosome formation.

1.9.1 Using DNA structural parameters to predict nucleosome positioning

The program BEND has often been used to predict DNA curvature and flexibility as a supplement to wet-lab mapping of positioned nucleosomes (Bash *et al.*, 2001; Blomquist *et al.*, 1999; Fiorini *et al.*, 2001; Wada-Kiyama *et al.*, 1999). The program accepts any DNA structural parameter set which can explain DNA bending along a DNA sequence, for example di-/tri- nucleotide parameter sets of twist, roll, tilt based on gel anomaly studies (Bolshoy *et al.*, 1991), cyclization kinetics (Ulanovsky *et al.*, 1986), X-ray crystallography (Calladine *et al.*, 1988) etc.. This software was useful to show that the binding of transcription factor *NF-1* depended on the position of curved DNA, which in turn affected nucleosome rotational positioning around the *NF-1* binding site (Blomquist *et al.*, 1999). The analysis was performed by introducing various sequence changes around the binding site and analyzing the potential effects of curvature. The software also helped to confirm bend sites, which were predicted using the circular permutation assay, in the promoter region of the GAL1-10 gene in yeast (Bash *et al.*, 2001).

The wavelet tool (used in this thesis; Section 2.4.1, Chapter 4) is an example of a different approach which can use DNA structural parameters. It can be used to assess the occurrence and distribution of structural patterns that could affect nucleosome positioning (Arneodo *et al.*, 1995; Arneodo *et al.*, 1998; Audit *et al.*,

2001; Audit *et al.*, 2002). So far, it has been used to show that non-coding eukaryotic genomic DNA contain periodic flexibility patterns (>100 bp periodic) which do not appear in coding DNA or in prokaryotic DNA sequences. The size of such repeat periods, which reflects the size of a nucleosome, has been suggested to be potential nucleosome-positioning elements.

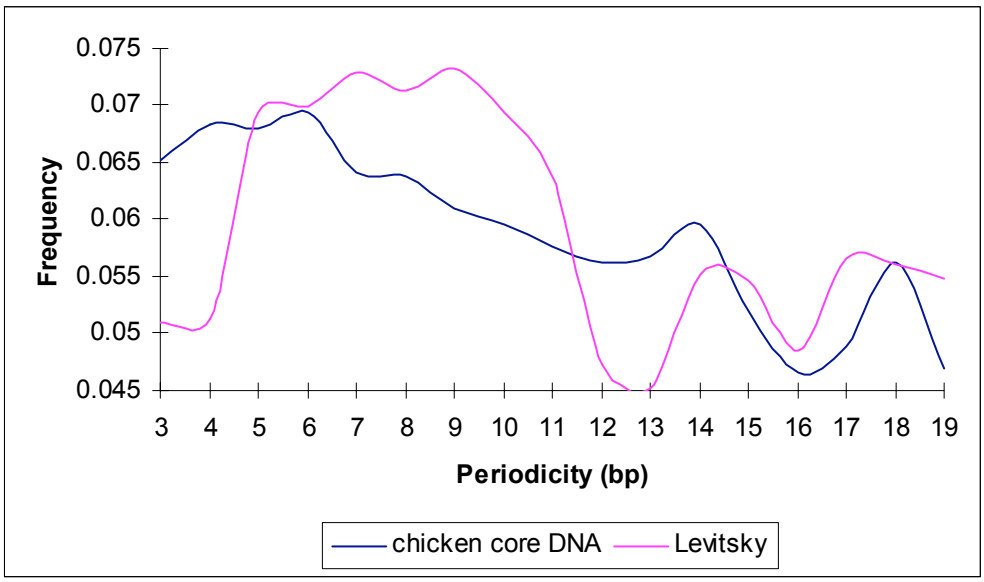
1.9.2 [AA/TT] rotational positioning pattern obtained using multiple sequence alignment

Ioshikhes *et al.* used five kinds of multiple alignment algorithms to create profiles of the nucleosomal database described earlier (Section 1.8.2) (Ioshikhes *et al.*, 1996; Ioshikhes & Trifonov, 1993). The algorithms considered only the positions of [AA/TT] dinucleotides because of their importance in rotational positioning described earlier (Section 1.4.1). These algorithms modelled an [AA/TT] dinucleotide positional frequency with a periodicity of 10.3(±0.2) bases towards the ends of a 146 bp sequence. [TT] dinucleotides also appeared to be distributed symmetrically relative to [AA] dinucleotides on the same DNA strand (phase difference: 6 bp). This result was reminiscent of the Fourier analysis results of the chicken core DNA dataset (Section 1.4.2) (Satchwell *et al.*, 1986) except the latter found [GC], rather than [TT], to be in phase with [AA]. A similarity, however, was that the periodic feature was seen to appear symmetrically away from the central 15 bp indicating that the DNA in the location of the dyad axis was not bent.

According to the multiple sequence alignment of these sequences using the software *Clustal W* (Appendix A), phased A-tracts were evident towards the first half of the sequences. However, the algorithms used to align the sequences by Ioshikhes *et al* were more strategic in that they did not model any 'deletes' and were specifically handling [AA/TT]-periodicity (*Clustal W* uses the 4-letter DNA alphabet and will

align any given sequences). Therefore, the alignment results from using *Clustal W* cannot be expected to give exactly the same results. Simple counting of [AA]-spacing showed a smeared peak between 5-11 bp for this dataset (Figure 1.5) indicating that phased-A tracts were featured in this dataset.

Figure 1.5: Simple counting of [AA]-spacing in the 2 experimentally-mapped nucleosome datasets (Section 1.8).



Denisov *et al.* used this model to predict nucleosome-centering around splice sites in 2000 exon-intron boundary sequences (400 bp fragments) obtained from a variety of eukaryotic species (Denisov *et al.*, 1997). The sequences appeared to position the midpoint of the nucleosome towards the introns. However, the data presented in the analysis were averaged values and it is not clear what proportion of the sequences showed this trend.

1.9.3 10-periodic [VWG] pattern obtained using hidden markov models

A 10-periodic [VWG] motif was found serendipitously using hidden markov models (HMMs) (Baldi *et al.*, 1996). Initially, conventional left-right hidden markov models, which were being trained to recognize splice-site junctions, learnt this signal. A

different kind of HMM architecture, the cyclical HMM was constructed which detected this motif with an apparent 10 bp periodicity in coding sequence. Many of the sequence members of the motif [VWG] were seen to be highly flexible in a DNase I – based flexibility table (Brukner *et al.*, 1995). This kind of proposed bending was different to the A-tract bending described earlier (Section 1.4.1); this suggests that 10-phased "flexible" motifs ([VWG]), rather than 10-phased "rigid" motifs ([AA]), could help to achieve nucleosome rotational positioning. The result was described as a flexible motif which appeared every 10 bp and which was superimposed over coding DNA[8]. This study suggested that exons could possess a nucleosome-binding signal superimposed over protein-coding signal.

Stein *et al.* used this observation as a model to predict nucleosome-positioning on the SV40 minichromosome simply by counting occurrences of 10-periodic [VWG] motifs (Stein & Bina, 1999). The results showed a weak correlation (correlation co-efficient: 0.52 with a P value <0.001) with experimentally-mapped nucleosomes in a 3,300 bp region (out of 5,200 bp) in the late SV40 region. It was described that in regions in the SV40 early region, where [VWG] could not be used to predict strong nucleosome positions, the 10-periodic [AA/TT] signal (Section 1.9.2) could. 5,000 bp is perhaps too short a sequence length for analysing nucleosome-positioning though: the maximum number of nucleosomes that could possibly fit on the whole SV40 minichromosome would be <30. Also, the reported correlation was observed in a specific part of the sequence rather than throughout the entire sequence.

[8] Coding DNA has harmonics of 3 bp.

1.9.4 RECON: A nucleosome prediction model based on dinucleotide relative abundance distance

A function to find 'nucleosome formation potential' was described recently (Levitsky *et al.*, 2001a). The prediction software, called RECON, was based on a function which calculated the optimal distance in dinucleotide space between mouse genome sequences that position nucleosomes (positive set) (Widlund *et al.*, 1997) and mouse genome sequences that repel nucleosomes (negative set) (Cao *et al.*, 1998). 86 sequences were available in the positive set and 40 sequences in the negative set. Using a jack-knifing procedure for model-testing, a model was trained which showed 80% accuracy at 94% coverage. Prediction analysis using this algorithm showed that introns and Alu repeats had a higher nucleosome formation potential than exons (Levitsky *et al.*, 2001b).

However, using fluorescence *in situ* hybridization, the positive set used in this study were found to belong to the mouse centromeric class of repeats (Widlund *et al.*, 1997). Centromeric nucleosomes are known to bind octamers, which have a variant of histone H3 in a large number of eukaryotes; this includes mouse (Smith, 2002). Therefore, it is unlikely that this positive set represents the majority of sequences that would bind nucleosomes in 'non-centromeric' genomic DNA.

The mouse positive sequences, used in RECON, were part of Levitsky *et al*'s nucleosome dataset introduced earlier (Section 1.8.2). However, the pair wise multiple sequence alignment of these sequences showed that a large number of the mouse sequences were highly similar to each other (Appendix A). These close variants were not reported to be discarded in the RECON software training. These could bias the results learnt in the RECON model.

1.10 Summary of Aims

The idea of nucleosome positioning, particularly its potential role in transcription regulation in eukaryotic cells, was an interesting prospect to research. With the large amount of eukaryotic genomic sequences now available from recent sequencing projects, particularly human and mouse data, an appealing option was to scan for evidence of nucleosome positioning, build models to predict nucleosome positioning and compare the predictions with known annotated features on these sequences.

1.10.1 The scope for studying nucleosome positioning

However, the scope for building good quality nucleosome models was limited. The restrictions arose partly from the limited experimentally-mapped data that supported nucleosome positioning. The 2 experimentally mapped nucleosome datasets (Section 1.8) each contained less than 200 sequences and also the initial sequence alignments of the 2 datasets did not show any obvious similarity between the 2 (Appendix A). About 36 sequences in the Levitsky dataset were also redundant.

Also, with regard to their role in events such as transcription regulation, the general view is that nucleosomes repress such activities (Section 1.4.4, 1.5.3); this could probably be a consequence of nucleosomes lying in the path of regulatory proteins such as RNA polymerase and transcription factors. This does not require nucleosomes to be positioned and it is not yet clear to what proportion positioned nucleosomes could repress transcription *in vivo*. Specific examples are available, for example *NF1*-binding to the MMTV promoter (Pina *et al.*, 1990) (Section 1.4.4). In this case, the position of a nucleosome is thought to be regulated by binding of a regulatory receptor protein, which in turn affects the accessibility of a transcription factor to its target site. From this, it could firstly be expected that it would not be

energetically favourable to have a large density of specifically positioned nucleosomes throughout the genome. Secondly, the few nucleosome positioning signals that are available could be expected to appear near gene regulatory regions where they could carry out important functional roles. Overall, this does make it difficult to detect nucleosome positioning sequences with high sensitivity especially from using whole genome analysis techniques.

The role of chromatin remodelling complexes (Section 1.4.4, 1.5.3) in directing nucleosome positions near promoter regions provides additional speculation that many nucleosomes could be positioned. In other words, it could be hypothesized that the remodelling complexes target positioned nucleosomes *in vivo*. At the moment, this remains speculation as the roles of chromatin remodelling complexes have not yet been assessed *in vivo* (Tsukiyama, 2002).

It is also important to note that the current experimental procedures used to reconstitute and map nucleosomes may not represent positioned nucleosomes *in vivo*. Chromatin extracts often contain much higher levels of the HMG (high mobility group) of chromatin proteins than the cellular background (Wolffe, 1998). These proteins are known to interact with nucleosomes. *In vivo*, chromatin structure is dynamic and using reconstitution procedures it is difficult to mimic the activity of important factors such as chromatin assembly factors, post-translational modification of histones and the nucleosome assembly process itself (which occurs in stages). Also, in the reconstitution procedure, it is quite difficult to assess the non-specific association of DNA with histones.

1.10.2 Aims and benefits of predicting nucleosome positioning

Given the limitations above, predicting nucleosome positioning was always going to be a challenging task. Most of the evidence for nucleosome positioning itself was

based on the results of *in vitro* experiments including the hypothesis of intrinsically curved DNA (Sections 1.4.1, 1.4.2). Possibly the major indication that nucleosomes could be positioned *in vivo* came from Lowary *et al*'s work, using competitive reconstitution (Section 1.7) (Lowary & Widom, 1997). From the results, it was estimated that only 5% of the mouse genome was probably enriched for binding nucleosomes

The aim in this thesis was to build computational models to predict nucleosome positioning. The first objective was to scan for evidence which could suggest that nucleosome positioning signals exist in the first place in eukaryotic genomic sequences. A second goal was to scan for evidence that suggests that nucleosome positioning could be involved in gene regulation. This would be carried out using 3 major modelling approaches (Section 1.11). If the positioning predictions, using any of the modelling techniques, indicated the following properties, it could suggest importance of nucleosome positioning in gene regulation *in vivo*:

- A high density of predictions in the vicinity of annotated genes
- Conservation of the prediction patterns in different eukaryotic species

If, however, the predictions were made randomly throughout the genome, it would suggest more that nucleosome positioning, if it does occur, is important only for maintaining and stabilizing higher order chromatin structures.

Being able to predict nucleosome positioning would definitely be beneficial in certain areas of genomic research. It may, for instance, aid in gene prediction if it can be shown that certain genes or regulatory DNA sequences have positioned nucleosomes over them or in their vicinity. This may, in turn, lead to clues about their expression patterns. Another area where it may be helpful is in the diagnostics of

chromatin diseases, many of which are postulated to be due to aberrant nucleosome positioning (Hendrich & Bickmore, 2001).

1.11 Approaches proposed for modelling nucleosome positioning

The methods outlined below have been employed in this thesis to approach the problem of predicting nucleosome positioning. Chapter 2 will give a brief summary of the theories of these methods.

1.11.1 Potential for studying 10 bp-phased motifs

Chapter 3 of this thesis deals with the use of cyclical HMMs. The aim of this approach was to scan for 10 bp-phasing motifs in genomic sequences, which could potentially influence nucleosome rotational positioning. This modelling approach extended the cyclical HMM work of Baldi and Brunak (Baldi *et al.*, 1996), which was introduced earlier (Section 1.9.3). The results obtained by Baldi and Brunak suggested that 10-phased [VWG] could be a nucleosome positioning signal. Many of the sequence members of this motif were highly flexible according to a DNase I–based flexibility table (Brukner *et al.*, 1995). Baldi and Brunak's overall technique, however, involved only learning the motif from various kinds of human genomic sequences including exons, introns and intergenic sequences: the models were not used to perform any predictions. The architecture of their cyclical HMMs was extended in this thesis to additionally model the background distribution of learnt 10-cyclical motifs. This would allow a HMM to be trained which could be used as a prediction tool. The two experimentally-mapped nucleosome datasets were also used as training sets for this purpose.

1.11.2 Potential for studying nucleosome translational positioning

In Chapter 4, the wavelet transform tool (Section 2.4.1) was used to probe the locations of periodic flexibility patterns in genomic sequences. The aim for the investigation was to establish whether any evidence existed suggesting that translational nucleosome positioning was an important mechanism for positioning nucleosomes in eukaryotic species. This would be achieved by modelling DNA sequences as flexibility sequences (Section 2.3.1). Recent work had already reported that eukaryotic DNA exhibit significant flexibility patterns which correspond to the repeat length of the nucleosome and which do not appear in prokaryotic genomes (Audit *et al.*, 2001; Audit *et al.*, 2002). It has also been reported that such patterns appeared only in non-coding DNA (Arneodo *et al.*, 1995; Buldyrev *et al.*, 1998; Havlin *et al.*, 1999; Pattini L, 2001). However, the genomic contexts of such patterns had not been clarified yet.

In Chapter 4, the wavelet transform tool was used to establish both the distribution of strong periodic flexibility patterns in representative genomes as well as determine if such patterns appeared near gene dense regions in DNA sequences. In addition to establishing the locations of these periodic features, it could also be determined if previously known DNA sequence features were the major players in determining potential nucleosome translational positioning.

1.11.3 Using DNA weight matrices to model the existing nucleosome datasets

The two available nucleosome datasets (Section 1.8) have both been analysed for rotational positioning and have been described to contain such positioning signals a

few turns away from and symmetrically about the nucleosome dyad axis (Ioshikhes *et al.*, 1996; Satchwell *et al.*, 1986) (Sections 1.4.2, 1.9.2). The methods applied themselves, however, were specifically aimed to find rotational positioning signals, namely patterns which recur at 10 bp periodicity in these datasets. For the chicken dataset, this was obtained using 3 bp window-averaged counts of dinucleotides along their position in the sequences (Satchwell *et al.*, 1986); this found the motif [AA/TT] to be enriched at 10 bp periodicity along with a relative 5 bp phase-shifted [GC/GC] motif. For Levitsky *et al*'s data, it was assumed that [AA/TT] was the major rotational positioning motif and the periodicity of this motif was analysed using several multiple sequence alignment algorithms (Ioshikhes *et al.*, 1996). This yielded a similar result to the chicken data except that [TT], and not [GC/GC], was reported to be phased at 5 bp to [AA] on the same strand.

However, to be a significant pattern, the suggested rotational positioning motifs should be present in the majority of these sequences; this has not yet been clarified for either dataset. Thus a motivation was formed to apply a rigorous classification system to each of the nucleosome datasets. This was the focus for the work in Chapter 5.

[RNA Biology 5:4, 181-188; October/November/December 2008]; ©2008 Landes Bioscience

Meeting Report

miRNA, siRNA, piRNA

Knowns of the unknown

Claudia Kutter[1,*] and Petr Svoboda[2,*]

[1]Cambridge Research Institute; Li Ka Shing Centre; Cambridge UK; [2]Institute of Molecular Genetics; Academy of Sciences of the Czech Republic; Prague Czech Republic

Key words: RNA silencing, miRNA, piRNA, siRNA, RNAi, non-coding RNA

RNA silencing is a common term for a group of mechanistically related pathways that produce and employ short non-coding RNA molecules to achieve sequence-specific gene regulation. The RNase III-enzyme Dicer produces small RNAs (smRNAs) in both microRNA (miRNA) and RNA interference (RNAi) pathways. miRNAs modulate physiological and developmental gene expression. They are genome-encoded, endogenous negative regulators of translation and mRNA stability originating from long primary transcripts with local hairpin structures. RNAi is triggered by the processing of long double-stranded RNA (dsRNA) into small interfering RNAs (siRNAs), which mediate sequence-specific cleavage of nascent mRNAs. The third common class of repressive small RNAs, PIWI-associated RNAs (piRNAs), is produced in a Dicer-independent manner. Current data suggest that piRNAs protect the germline from mobile genome invaders such as transposons. A small RNA involved in RNA silencing associates with proteins in an effector ribonucleoprotein complex usually referred to as RNA-Induced Silencing Complex (RISC). Key components of RISC complexes are proteins of the Argonaute family, which determine RISC functions. During three days in May 2008, around two hundred scientists working on RNA silencing met at IMBA, Vienna for the 3rd Microsymposium on Small RNAs (Fig. 1) (www.imba.oeaw.ac.at/microsymposium) organized by Javier Martinez.

RNA Silencing in Plants

Small RNAs in plants play critical roles in developmental, defence and stress processes, and perturbation of their function can have dramatic consequences. A considerable effort is now given towards understanding how smRNA steady state levels are regulated and maintained. Three talks were devoted to miRNAs (Fig. 2A) and related silencing pathways in plants.

First, **Allison Mallory** (INRA; Versailles, France) reviewed RNA silencing pathways in plants and then discussed results of a forward genetic suppressor screen to identify genes whose mutations suppress *ago1* mutations and restore RNA silencing. This screen identified FIERY1 (FRY1) as an endogenous posttranscriptional gene silencing (PTGS) suppressor in Arabidopsis. Reverse genetics approaches identified *XRN2*, *XRN3* and *XRN4* as targets of FRY1 and revealed that these 5'–3' exoribonucleases also function as endogenous PTGS suppressors. The absence of the nucleotidase/phosphatase FRY1 causes accumulation of miRNA cleavage products and excised miRNA loops, which are templates of the cytoplasmic XRN4 and nuclear XRN2 and XRN3, respectively. Consistent with their role as PTGS suppressors, *fry1* and *xrn4* mutants are hyperresistant to cucumber mosaic virus infection. In addition, mutations in these suppressors revealed that XRN3 is essential for plant viability.[1]

The next talk by **Javier Palatnik** (Institute of Molecular and Cell Biology of Rosario, Argentina) demonstrated a cascade function of the miR159/miR319 family in Arabidopsis during leaf development. The miR159/miR319 system represents an interesting model for studying functional specialization of highly related miRNAs. These miRNAs are highly similar in sequence but specifically regulate either *MYB* or *TCP* transcription factors, which quantitatively regulate cell proliferation. This study of interactions of miRNAs with nascent mRNAs revealed that the degree of interaction of different miRNAs with the same site is a key feature for quantitative regulation of gene expression. Improved basepairing of miR159a by changing three nucleotides (nt) in the target mRNA (*TCP*) results in efficient repression of *TCP* while overexpression of wild-type miR159 has no effect on *TCP*. In leaves, adequate control of cell proliferation is achieved through the specific regulation of *TCP* gene expression by miR319. TCP is a repressor of cyclin genes and thus cell proliferation. Interestingly, miR319 interacts with miR396, another miRNA involved in leaf development in plants.

The third talk on RNA silencing in plants was given by **Xuemei Chen** (University of California; Riverside, CA USA) who discussed small RNA metabolism. Plant small RNAs are modified by the methyltransferase HEN1, which adds 2'-O-methyl group at the 3' terminal nucleotide. This modification has a stabilizing effect on small RNAs. Small RNAs from *hen1* mutants exhibit heterogeneity at their 3' end, which seems to be caused by uridinylation and exonucleolytic degradation. Turnover of miRNAs was further addressed by the identification of SMALL RNA DEGRADING NUCLEASE 1 (SDN1), an exonuclease, that degrades single-stranded small RNAs

*Correspondence to: Claudia Kutter; Cambridge Research Institute; Li Ka Shing Centre; Robinson Way; Cambridge CB2 0RE UK; Tel.: +44.1223.404501; Fax: +44.1223.404199; Email: Claudia.Kutter@cancer.org.uk / Petr Svoboda; Institute of Molecular Genetics; Videnska 1083; 14220 Praha 4, Czech Republic; Tel: +42.0 2.4106.3147; Fax: +42.02.2431.0955; Email: svobodap@img.cas.cz

Submitted: 05/29/08; Accepted: 10/16/08

Previously published online as an *RNA Biology* E-publication:
http://www.landesbioscience.com/journals/rnabiology/article/7227

Figure 1. Logo of IMBA, Vienna.

in vitro. Knock-down of the exonuclease gene in vivo results in elevated miRNA levels.[2]

miRNAs and miRNA Targets in Animals

Analysis of the miRNA pathway in animals (Fig. 2B) was the major theme of the conference. As a result, it was covered in approximately half of the talks. Analysis of physiological functions of miRNAs in different animal model systems is booming and an apparent trend is combining biological data with extensive in silico analysis. Whereas predictions based solely on sequence analysis are typically insufficient for understanding miRNA function in a specific model system, the combination of appropriate global profiling data with bioinformatic analysis is more powerful and often leads to improved understanding of miRNA roles. In any case, reliable identification of miRNA targets remains one of the key objectives in the miRNA field because distinguishing between real targets and false positive ones is a persistent problem.

Uwe Ohler (Duke University; Durham, NC USA) discussed the distribution of miRNA target sites in 3'UTRs and computational methods to improve miRNA target prediction. He introduced a concept of alternative polyadenylation (polyA), a common feature of many genes, which use more than one polyA signal to define the 3' end of mRNA. The 3'UTR defined by the first polyA site is called constitutive 3'UTR while the UTR defined by a later polyA site is called alternative. Alternative polyadenylation thus offers another layer of control of gene expression because additional sequences in the 3'UTR allow regulation of a specific subset of transcripts coding for the same protein. Ohler showed that approximately 60% of predicted miRNA targets are in genes with alternative 3'UTRs, with 40% of predicted miRNA binding sites located in alternative UTRs. Target site distribution along the 3'UTR is nonrandom as target sites appear to cluster around polyA sites.[3] Ohler also discussed identification of enriched motifs in sets of 3'UTRs. Analysis of enriched motifs in 3'UTRs is an increasingly popular strategy for data mining microarray profiles from cells with a perturbed miRNA pathway. Ohler used microarray data from cells overexpressing a miRNA to test a new motif-finding algorithm. This new approach allows identifying degenerate motifs. To improve efficiency of identification of enriched miRNA binding sites and to look beyond the seed motif, he included a folding approach. Using the new set of rules, he achieved a two-fold enrichment on sites with 6 to 8 nt matches in the seed sequence stressing the importance of site accessibility in miRNA target recognition.

A talk by **Joel Neilson** (Massachusetts Institute of Technology; Cambridge, MA USA) resonated with Uwe Ohler's discussion of alternative usage of polyA sites. Neilson presented data suggesting that T-cells tend to express transcripts derived from the use of upstream polyA sites following activation. He used exon array profiling to obtain comprehensive information about alternative splicing and alternative usage of polyA sites. Neilson proposed a general association of the shift to shorter 3'UTRs with higher proliferation.[4] Furthermore, he suggested that, since mRNAs species with short 3'UTRs typically contain half the number of target sites compared to those with long 3'UTRs, this might explain differences in miRNA-mediated gene regulation in undifferentiated cultured cells versus differentiated tissues and could be one of important mechanisms dysregulated in cancer.

The extent of miRNA function in animal cells has largely been studied by mRNA microarray profiling assuming that miRNA function leads to reduced mRNA levels, which may not be always the case. In a pioneering work, **Nikolaus Rajewsky** (Max Delbrück Center for Molecular Medicine; Berlin, Germany) presented a study of genome-wide regulation of translation by miRNAs based on proteomic analysis by pSILAC, a modern mass-spectrometry method that was adapted to measure changes of newly synthesized proteins.[5] As a model, Rajewsky and colleagues chose HeLa cells transfected with five different miRNAs. pSILAC analysis showed very good reproducibility and turned out to be a good method for identifying miRNA-mediated translational repression. Eight miRNA targets were validated using dual luciferase reporters. This correlated well with pSILAC data. Motif analysis showed enrichment of pSILAC data with miRNA targets. About 60% of

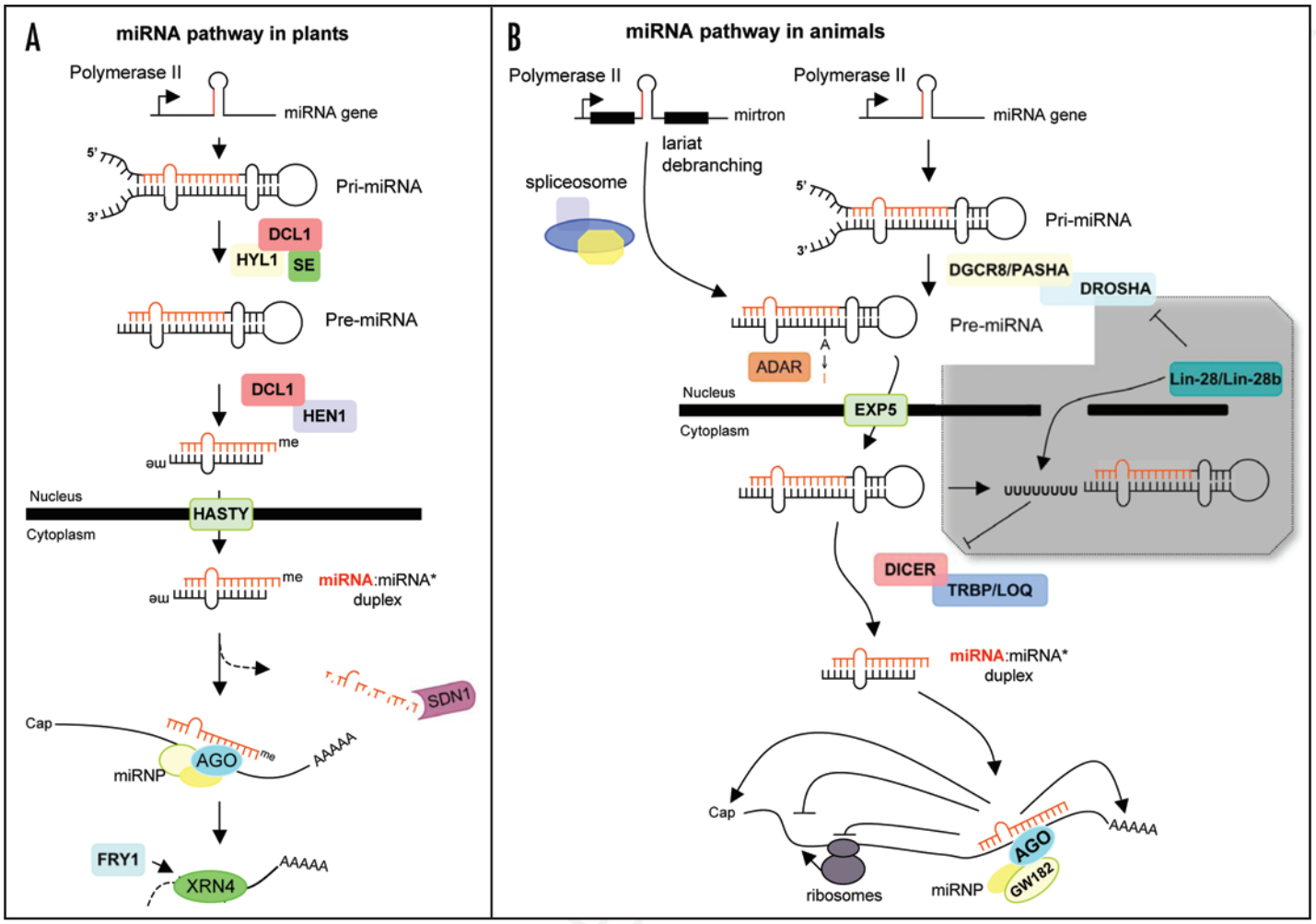


Figure 2. Mechanism of microRNA (miRNA) biogenesis and miRNA-mediated mRNA silencing in plants and animals. (A) plant miRNAs are generated by a DICER-LIKE 1 (DCL1)-mediated cleavage of RNA polymerase II-transcribed endogenous precursor. HYPERNASTIC LEAVES 1 (HYL1) and SERRATE (SE) are involved in the processing of the precursors to form a miRNA duplex. Each 3' end of the duplex is protected from polyuridylation and degradation by a methyl group added by the methyltransferase HUA ENHANCER 1 (HEN1). The miRNA duplex is shuttled to the cytoplasm by HASTY (HST). In there, the steady-state level of miRNAs is controlled by the exonuclease SMALL RNA DEGRADING NUCLEASE 1 (SDN1). The guiding strand (miRNA) of the miRNA duplex is incorporated into an ARGONAUTE 1 (AGO1) containing protein complex (miRNP). The miRNA guides RISC to the target mRNA leading to mRNA degradation. FIERY1 (FRY1) controls EXORIBONUCLEASE 4 (XRN4) activity, which degrades the 3' fragment of miRNA-cleaved mRNAs. (B) miRNA transcripts in animals adopt an imperfect fold back structure and undergo sequential processing by the microprocessor (Drosha and Pasha in flies; Drosha and DiGeorge syndrome region gene 8 protein (DGCR8) in mammals) to form precursor (pre-miRNA) intermediates. Mirtrons are short intronic hairpins that are spliced and debranched to mimic pre-miRNAs. In some cases single nucleotides of the pre-miRNA are modified by adenosine deaminase (ADAR). Both types of pre-miRNAs are exported to the cytoplasm via Exportin-5 (EXP5) where they are processed by Dicer (in complex with Loquacious (LOQS) in flies or trans-activation responsive RNA-binding protein (TRBP) in mammals). The grey box highlights inhibition of miRNA processing by Lin-28/Lin-28b that is so far only specific to let-7 family members. On one hand, Lin-28/Lin-28b may prevent access of Drosha (and DGCR8) by binding to the loop region of pri-/pre-miR-let-7 in the nucleus. On the other hand, Lin-28 may block Dicer-processing in the cytoplasm by facilitating uridylation and therefore degradation of pre-let-7a. miRNAs are assembled in ribonucleoprotein complexes (miRNPs). P-body components such as GW182 interact with Argonaute (AGO) in miRNPs to either promote inhibition of translational initiation at the cap-recognition stage or the joining of the 40S and 60S subunits stage. As alternatives miRNPs can accelerate deadenylation and decay of target mRNAs, block off elongation or ribosome "drop-off," or cause degradation of nascent polypeptides.

proteins downregulated by at least 20% could be explained by the presence of a seed motif and conserved seeds appeared to have a stronger impact on protein downregulation than unconserved seed motifs. By performing transcript profiling in parallel and subtracting this from the measured changes in protein synthesis they could show that miRNAs directly repress translation of hundreds of target genes. Knockdown of an endogenous miRNA validated these results and showed that a single miRNA can tune protein levels of thousands of genes.

Current conditional knock-outs of individual miRNA pathway components allow dissection of the roles they play in specific processes in vivo. **Donal O'Carroll** (European Molecular Biology Laboratory; Monterotondo, Italy) studied the physiological role of RISC complexes in hematopoiesis using a conditional knock-out mouse model. The MxCre-induced loss of Ago2 caused defects in B-cell and erythroid lineages. Surprisingly, slicer inactive Ago2 bearing D699A mutation in the PIWI domain rescued the Ago2$^{-/-}$ phenotype demonstrating that the slicer activity is dispensable for

Ago2 function.[6] Remarkably, Ago2-containing RISC is the functionally dominant RISC variant because loss of Ago1 and Ago3 (which are expressed in bone marrow and have no "slicer" activity) has no effect on hematopoiesis. These observations further support a general notion that, while all Ago proteins are associated with miRNAs and act in the miRNA pathway, Ago2 is functionally dominant and the contribution of other Agos to miRNA-mediated silencing is non-essential. At the same time, the loss of Dicer arrests hematopoiesis earlier and affects also T-cell development, suggesting that either Ago1 and Ago3 delay the onset of $Ago2^{-/-}$ phenotype and/or Dicer has some miRNA-independent role. Finally, O'Carroll pointed out that Ago2 acts upstream in the processing/stabilization of pre-miRNAs, but Ago1 and Ago3 do not.

A rare insight into miRNA function in an adult organism was offered by **Klaus Förstemann** (Ludwig-Maximilians University of Munich; Munich, Germany) who addressed the biology of miR-277 in Drosophila. He showed that miR-277 is preferentially loaded into the typically siRNA-bearing AGO2-containing RISC complex. Target prediction (www.russell.embl.de/miRNAs/) yielded over 400 putative targets highly enriched in metabolic genes, particularly of mitochondrial metabolism and aminoacid degradation. Localization studies showed low levels of miR-277 expression in larva and pupa and high levels in adults. It was speculated that miR-277 might act as a metabolic switch. Interestingly, overexpression of miR-277 is lethal when the caloric content of the food is low and this phenotype can be rescued by providing sugar but not by providing proteins. It has been proposed that miR-277 might target genes encoding for enzymes that control the conversion of amino acids into intermediates of the citric acid cycle, thus explaining the particular dependence on sugar in the food when miR-277 is overexpressed.

Leonard Goldstein (Simon Tavaré lab, Cambridge Research Institute; Cambridge, UK) presented data on miRNA and mRNA expression during mouse mammary gland development. He was investigating whether breast cancer associated miRNAs[7] play a role in normal mammary gland development. He and colleagues profiled whole mammary glands at 15 developmental time points including lactating glands, in which he observed reduced miRNA levels. Altogether, he identified approximately 120 miRNAs expressed at different stages of mammary gland development, with most miRNAs following one of eight typical expression profiles. Members of the same miRNA family frequently showed highly correlated expression, as did orthologs of miRNAs associated with human luminal breast cancers.

Let-7 and AAGUGC-miRNAs—First Among Equals

Two miRNA families dominated presentations involving specific miRNAs: let-7 and miRNAs bearing the AAGUGC motif at their 5' terminus. Let-7 is the favorite model for studies of miRNA-related mechanisms, which is not surprising considering let-7 is one of the longest-studied miRNAs and it is highly conserved. **Amy Pasquinelli** (University of California, San Diego, CA USA) focused on the regulation of miRNA expression in *let-7* and *lin-4* models. At least two different isoforms of pri-miRNA *let-7* exist in *C. elegans* due to different transcription start sites. In addition to transcriptional regulation, post-transcriptional mechanisms may also contribute to restriction of *let-7* miRNA to the appropriate time in development since a GFP reporter gene fused to the *let-7* promoter is detectable earlier than endogenous miRNA expression. Cis-acting modulators might also regulate the transcriptional switch of two isoforms of *lin-4* pri-miRNA during development with only one form apparently entering the miRNA biogenesis pathway.

Xavier Ding (Helge Grosshans lab, Friedrich Miescher Institute; Basel, Switzerland) discussed how depletion of translation factors suppresses *let-7* lethality in *C. elegans*. An RNAi screen of 2400 genes aimed to identify interaction partners of *let-7* identified several translation initiation factors as suppressors of the *let-7* lethality phenotype. Consistent with this finding, polysome profiling of miRNA targets in various genetic backgrounds revealed that miRNAs regulate target genes in vivo at the level of translational initiation. Furthermore, Ding's data also suggest that miRNA-mediated target mRNA degradation frequently occurs together with translational repression but functions as an independent mechanism.

Inha Heo (Narry Kim lab, Seoul National University; Seoul, South Korea) showed that Lin-28 inhibits let-7 biogenesis after Drosha processing. Lin-28 associates predominantly with pre-let-7a in the cytoplasm as shown by immunoprecipitation. In vitro studies demonstrated that Lin-28 mediates uridylation of pre-let-7a and promotes its degradation. The U-tail irreversibly blocks Dicer-processing of pre-let-7a. These findings lead to the conclusion that Lin-28 associates with pre-let-7a and recruits uridylating enzymes. **Inha Heo's** data partially overlapped with data presented by **Scott Hammond** (University of North Carolina; Chapel Hill, NC USA) who used a Let-7 processing model system to search for negative regulators of miRNAs. Let-7g is ubiquitously expressed but the mature miRNA is not found in early developmental stages while pri-miRNA is readily detectable. Using Let-7 loop sequence as a decoy, his group identified Lin-28 as a candidate inhibitor of Let-7g processing. Lin-28 is necessary and sufficient for Let-7g repression, which occurs at the Drosha processing step and involves Lin-28 interaction with Let-7 loop. Uridylation was not reported in this case (Fig. 2B).[8]

Scott Hammond also discussed the role of miRNAs in cancer and correlated several patterns of miRNA expression during development and cell differentiation with patterns during carcinogenesis. Many miRNAs, including Let-7, have reduced expression in various cancers. Hammond proposed that pathways leading to reduce miRNA expression might be important factors during tumorigenesis. Restoration of such pathways should have an anti-cancer effect. He used a Let-7 processing model system to study negative regulators of miRNAs and identified several candidate inhibitors of Let-7g processing. One of them, Lin-28, was characterized further, revealing that Lin-28 is necessary and sufficient for Let-7g repression. Inhibition of let-7 in early embryos of Drosophila, which is abolished in later developmental stages, points towards Lin-28 as a regulator of pluripotency and Let-7 miRNA as a "differentiation" factor. An analogous mechanism can be envisioned for tumor cells (HeLa cells), in which high levels of Lin-28b also suppress the expression of let-7 to promote growth.

The proposed role of Let-7 in negative regulation of "stemness" by repressing self-renewal and promoting differentiation in both normal development and cancer[9] places Let-7 opposite to the second largely discussed group of miRNAs: AAGUGC seed-containing miRNAs. AAGUGC seed is a prominent motif in several miRNA

families expressed in zebrafish (miR-430 family[10]) and mouse zygotes (miR-290 family[11]), in mouse and human embryonic stem (ES) cells (miR-290 family and related miRNAs[12,13]), and also in miRNAs with oncogenic potential (oncomirs, miR-17-92 cluster[14]).

Andrea Ventura (Memorial Sloan-Kettering Cancer Center, New York, NY USA) provided interesting insights into redundancy and specificity of miRNAs by presenting a study of deficiencies in the miR-17~92 cluster and its two paralogous cluster miR-106b~25 and miR-106a~363. Mice deficient for miR-17~92 cluster (also know as oncomiR-1) died shortly after birth with lung hypoplasia and a ventricular septal defect whereas mice lacking the paralogous cluster miR-106b~25, miR-106a~363, or both had no obvious phenotypic abnormalities.[15] A more detailed analysis revealed possible causes of different phenotypes. The miR-106a~363 expression was never observed. The miR-106b~25 cluster is expressed but lacks miRNAs from some of the miRNA families represented in the oncomiR-1 cluster. The miR-106b~25 cluster is partially compensating the loss of the oncomiR-1 cluster and the double knock-out causes an even more severe phenotype. The animals die at mid-gestation and the phenotype is associated with enhanced apoptosis. OncomiR-1$^{-/-}$ mice also show defects in B-cell development due to increased apoptosis of B-cell progenitors. The phenotype was also tested in adult mice using fetal liver reconstitution and a conditional miR-17~92 knock-out allele. Search for putative candidate miRNA targets whose misregulation would contribute to the defective B-cell development identified Bim, a tumor suppressor, which is a critical regulator of B-cell survival.

Lasse Sinkkonen (Witold Filipowicz lab, Friedrich Miescher Institute; Basel, Switzerland) discussed the role of miRNAs in ES cells and during their differen-tiation. Microarray analysis of Dicer$^{-/-}$ ES cells revealed that 3'UTRs of upregulated transcripts show highly significant enrichment of motifs complementary to AAGUGC, a seed found in ES-specific miRNAs of the mir-290 cluster and other miRNAs. Consistently, many of the defects in Dicer-deficient cells were reversed by transfection of miR-290 family miRNAs. One of such defects of Dicer$^{-/-}$ ES cells was the loss of de novo DNA methylation during differentiation, which correlated with downregulation of de novo DNA methyltransferases (Dnmts) in undifferentiated Dicer$^{-/-}$ ES cells. The downregulation was mediated by a potential direct miR-290 cluster target Rbl2. The defective DNA methylation could be rescued by ectopic expression of de novo Dnmts or by transfection of the miR-290 cluster miRNAs, indicating that de novo DNA methylation in ES cells is controlled by miRNAs.[16]

Petr Svoboda (Institute of Molecular Genetics; Prague, Czech Republic) discussed several aspects of maternal mRNA degradation in the mouse. Microarray analysis of maternal mRNA degradation suggests that miRNAs play a limited role and do not generate any detectable signature in the 3'UTR composition of the downregulated maternal transcriptome. However, about 20% of putative miR-290 targets identified in ES cells exhibit high expression in the oocyte when compared to the blastocyst and somatic tissues. As transcripts targeted by the miR-290 cluster in ES cells are not maternal, it has been proposed that an important role of zygotic miRNAs is to restrict zygotic expression of genes, which were highly expressed in the oocyte, while degradation of true maternal transcripts would be less important.

This proposed role of the miR-290 cluster in the mouse zygote is somewhat different from the zebrafish model, which was discussed by **Antonio Giraldez** (Yale University; New Haven, CT USA). Previous studies of the role of the zygotically expressed miR-430 family (closely related to the miR-290 cluster) in maternal Dicer null mutants (MZ Dicer) suggested that miR-430 is involved in degradation of up to 40% of maternal mRNAs.[10] Despite these differences, it is apparent that mouse and zebrafish employ related miRNAs in reprogramming of the zygotic genome. To further investigate miR-430 function, morpholino antisense oligonucleotides were used that are complementary to mRNA target sites and block miRNA-mediated repression of specific targets.[17] This approach is likely to become a valuable tool for analysis of specific miRNA-mRNA interactions. It was shown that miR-430 regulates Nodal signalling by balancing the expression of an agonist (squint) and an antagonist (lefty), genes involved in mesoderm expansion and reduction, respectively. The remainder of the talk focused on the role of tissue-specific miRNAs miR-1, miR-206 and miR-133. Morpholino oligonucleotides were used to identify miRNA-regulated target genes involved in myogenesis by microarray profiling and analysis of genes upregulated upon morpholino delivery.

Nuclear Aspects of RNA Silencing

Several talks touched upon nuclear aspects of RNA silencing, which have been considerably dissected in plants and *S. pombe* but are poorly understood in animals. **Marc Bühler** (Friedrich Miescher Institute; Basel, Switzerland) presented the function of Ago1 and Cid14 poly(A) polymerase in heterochromatic gene silencing mechanisms in *S. pombe*. Efficient silencing of some heterochromatic regions in fission yeast requires the RNAi machinery as well as the TRAMP complex, which contains the non-canonical poly(A) polymerase Cid14 and targets aberrant RNAs for degradation by the exosome.[18] Deep sequencing showed that, in wild-type cells, most Argonaute-associated small RNAs (smRNAs) start with a 5' uracil, are ~22/23 nt long, and match to repeated regions of the genome. In cells lacking Cid14, the small RNA profile changes include major new classes of small RNAs that originate from ribosomal RNAs and tRNAs. Bühler suggested that Cid14 prevents certain abundant RNAs from becoming substrates for the RNAi machinery, thereby freeing the RNAi machinery to act on its proper targets and to prevent spurious generation of siRNAs.[19]

While small RNAs can induce transcriptional silencing in mammalian cells, the mechanism of silencing is unknown. Data from **Lasse Sinkkonen**, who demonstrated that defects in de novo DNA methylation in Dicer$^{-/-}$ ES cells can be explained by defects in miRNA-dependent regulation of gene expression (reviewed in more detail above), implies caution as warranted when interpreting nuclear effects of RNA silencing in mammalian cells because they could be indirectly mediated by miRNAs.

Mariano Allo (Alberto Kornblihtt lab, University of Buenos Aires; Buenos Aires, Argentina) presented data on siRNA-mediated epigenetic control of alternative RNA splicing. He showed that siRNAs directed against intronic sequence could affect alternative splicing in human cells. The phenomenon was dependent on AGO1. It has been speculated that AGO1 might recruit the transcriptional silencing machinery, which affects chromatin in the targeted locus. This is consistent with siRNA-induced dimethylation at the target

site, which could cause pausing of polymerase II and, in turn, alternative exon inclusion. Likewise, the siRNA effect on alternative splicing is abolished by a specific inhibitor of H3K9 methyltransferase. However, the exact mechanism of siRNA-induced alternative splicing is not known at present.

One of the unknowns in nuclear RNA silencing in mammals is the "nuclear life" of RNA silencing components. An interesting contribution to this topic was presented by **Thomas Ohrt** (Ph.D. student award, Petra Schwille lab, Dresden University of Technology; Dresden, Germany), who investigated the nuclear role of AGO2 using fluorescence correlation spectroscopy (FCS) and fluorescence cross-correlation spectroscopy (FCCS). These methods allow detailed analysis of nuclear and cytoplasmic RISC complexes with single molecule sensitivity in vivo. Ohrt demonstrated that two different RISC complexes exist; one is present in the cytoplasm and the other in the nucleus. The nuclear RISC is much smaller and likely composed of just AGO2 and siRNA. In the nucleus, RISC can bind and cleave its targets. In the absence of cleavage, RISC accumulates in the nucleus in a target-dependent manner. FCS and FCCS methods hold great promises for the future and hopefully will bring new insights into nuclear aspects of RNA silencing.

piRNAs and Other Small RNAs

The world of small RNAs includes other classes that belong to the complex of related pathways termed RNA silencing. The second most discussed class of small RNAs were piRNAs, short single-stranded RNAs arising from a Dicer-independent pathway, which are found in germ cells and associate with the PIWI subfamily of Argonaute proteins.

Saskia Houwing (Ph.D. student award, Rene Ketting lab, Hubrecht Institute; Utrecht, The Netherlands) studied the role of Ziwi[20] and Zili (Piwi homologues) in the zebrafish germline. piRNAs isolated from zebrafish testes and ovaries resemble piRNAs from other organisms and this is suggestive of the presence of a ping-pong model for piRNA amplification in vertebrates. Many zebrafish piRNAs are derived from repetitive sequences. Mutations in the Zili protein resulted either in loss of germ cells or in defects in meiosis and chromosome segregation in eggs. Requirement of piRNA pathway in the female germline is in interesting contrast to the situation in mammals, where piRNAs are non-essential in the female germline but required for proper sperm development. Furthermore, defects in chromosome segregation in eggs were also observed in Dicer knockout mouse eggs[11,21] raising a question about overlapping functions of vertebrate miRNA, RNAi and piRNA pathways.

Alexei Aravin (Greg Hannon lab, Cold Spring Harbor, NY USA) presented the role of mouse Piwi homologues Mili,[22] and Miwi2 in the male germline and showed that repeat-associated piRNAs are present throughout spermatogenesis. Data reminiscent of the ping-pong model proposed for piRNA biogenesis in Drosophila showed that piRNAs are derived from two repeat elements: LINE L1 and SINE B1. A strand bias of those piRNAs was observed. One of the processing enzymes Miwi2 associates primarily with piRNAs deriving from the antisense strand whereas the other enzyme Mili associates with piRNAs sense strand. Interestingly, strand bias switching is observed at different developmental stages. Aravin also reported an additional class of endogenous siRNAs from mouse ovaries, which are 20–23 nt long and derive from double-stranded RNAs formed by pairing of spliced transcripts from protein-coding genes to antisense transcripts from homologous pseudogenes in a Dicer-dependent manner. Interestingly, their targets are mostly found in the microtubule/spindle pathway.[23]

Michael Reuter (Ramesh Pillai lab, European Molecular Biology Laboratory; Genoble, France) presented localization and interaction studies of the murine piRNA methyltransferase mHen1. Immunostaining revealed that mHEN1 is present in single cytoplasmic foci in round spermatids. These structures, named Hen bodies, differ from chromatoid bodies, which harbor Piwi proteins. Although co-immunostaining with MILI was undetectable, an interaction of endogenous proteins in testis extracts, and of HA-tagged mHEN1 with His-tagged MILI in HeLa cells, could be shown by co-immunoprecipitation. RNaseA treatment of the mHEN1-MILI complex showed that this interaction is RNA-independent. These findings uncover a new cytoplasmic "spot" that might participate in the final step of piRNA biogenesis.

Two new classes of mammalian small RNAs were presented during the meeting. **Christine Ender** (Gunter Meister lab, Max Planck Institute of Biochemistry; Martinsried, Germany) talked about 21–22 nt small RNAs derived from H/ACA-box small nucleolar RNAs (snoRNAs). These smRNAs derive from a specific region in the 3'arm of the snoRNA. Their biogenesis is Drosha and most likely Dicer-independent. Computational analysis indicates regulatory functions for these small RNAs in degrading or repressing target genes by binding to 3'UTRs.

Kata Fejes-Toth (Greg Hannon lab, Cold Spring Harbor, USA) presented a study of another class of small RNAs of unknown size that were identified by high-throughput screening. These promoter-associated small RNAs (PASRs) have an unknown length but cover a region of about -500 to +500 bp relative to transcriptional start sites. They apparently differ from small RNAs implicated in RNA silencing because conventional miRNA cloning procedures failed to identify them. The PASR population differs in HeLa and HepG2 cells and possibly only a small fraction has been cloned so far. PASRs are typically found within 0.5 kb of transcription start sites and about 40% of them map to 5' expressed sequence tags. Their function is unknown so far but they possibly regulate transcription as exogenous PASRs can reduce expression of genes with homologous promoter sequences.

Unique small RNAs, unrelated to small RNAs of RNA silencing pathways, also regulate gene expression in bacteria. The vast diversity of small bacterial RNAs was demonstrated **Jörg Vogel** (MPI for Infection Biology, Berlin, Germany), who reported the identification and the characterization of more than 25 *Salmonella enterica* small RNAs (sRNA) and their targets. These RNAs are usually 50 to 150 nt long and often act as antisense RNAs. He discussed in detail the role of GcVB RNA in the regulation of ABC transporters. The 200 nt long GcVB sRNA, which is expressed in fast growing cells, binds via a particular seed sequence to C/A-rich sites of mRNAs of ABC transporters; these sites serve as both translational enhancer regions and as GcvB RNA binding sites.[24] Toeprinting assays showed that GcVB likely mediates translational repression by inhibiting translation initiation. Another small RNA in Salmonella is the SigmaE-controlled RybB RNA, which is a multiple target regulator that uses a seed-like pairing similar to eukaryotic miRNAs.[25] It binds to 5' UTRs where it affects translation by inhibition of

30S complex binding, destabilization of 5' protective hairpin or by inhibiting ribosome elongation.

Technology

Gregor Obernosterer (Javier Martinez lab, Institute of Molecular Biotechnology; Vienna, Austria) discussed criteria of siRNA design to improve RNAi efficiency. Inspired by previous biochemical studies of Stefan Ameres on RISC-target interactions, Obernosterer and colleagues tested whether target site accessibility, as modeled in silico, affects siRNA efficiency. They found that the asymmetry of the siRNA complex and the accessibility of the target site for siRNA interaction are the most important criteria for efficient siRNAs whereas free self-folding and G/C content are less significant criteria. These findings were used in the development of a novel siRNA design tool, called RNAxs (http://rna.tbi.univie.ac.at/cgi-bin/RNAxs).[26] RNAxs ushers in a new generation of siRNA design tools, which also evaluate siRNA interaction with its target site. Accessibility also seems to affect miRNA targeting efficiency and might be a useful feature to incorporate into miRNA target prediction analysis.

A number of technologies and instrumental platforms for RNAi and miRNA studies were discribed by company representatives. **George Tzertzinis** (New England Biolabs; Ipswich, MA USA) discussed the use of complex siRNA mixtures. **Dmitry Samarsky** (RXi Pharmaceuticals; Worcester, MA USA) presented development of a novel RNAi therapeutic platform. Delivery and use of RNAi was a common theme in the presentations of **Jon Karpilow** (Thermo Fisher Dharmacon; Lafayette, CO USA), **Ludger Altrogge** (Amaxa; Gaithersburg, MD USA), **Francisco Bizouarn** (BioRad; Hercules, CA USA) and **Jens Harborth** (CequentPharma; Cambridge, MA USA). **Kai Wilkens** (Panomics Central Europe) advertised bDNA technology, a system for profiling gene expression. Analysis of miRNAs and other small RNAs using microarray platforms was presented by **Sonja Vorwerk** (Febit; Heidelberg, Germany), **Hazel Pinheiro** (Exiqon; Vedbaek, Denmark) and **Christoph Eicken** (LC Sciences; Houston, TX). Quantification and functional analysis of miRNAs was discussed by **Constanze Kindler** (Qiagen; Cologne, Germany). Particular attention was brought to locked nucleic acid (LNA) technology in a workshop organized by Exiqon. Due to their thermodynamic properties, LNA oligonucleotides seem to be excellent reagents for hybridization based assays. **Timothy Gant** (University of Leicester; Leicester, UK) and **Daniele Catalucci** (ITB-CNR; Milan, Italy) presented the use of LNA oligonucleotides in analysis of miRNAs. Technical details about using LNA probes for in situ hybridization were discussed by **Gregor Obernosterer** (Javier Martinez lab, Institute of Molecular Biotechnology; Vienna, Austria).[27]

Summary and Outlook

Meetings like this are rare—no registration fee, easy-to-approach attendees presenting the latest research, an attractive location, smooth organization and great hospitality. The Microsymposium in Vienna is a stimulating and inspiring event for Ph.D. students, postdoctoral researchers and group leaders. In fact, the meeting encourages and provides the opportunity for young scientist to present their work as we could see in the outstanding Ph.D. student presentations and a high-quality workshop on miRNA analysis organized by Exiqon.

The 2008 Microsymposium meeting highlighted the importance of small RNAs in gene regulation, their use in research and potential application in diagnostics and therapeutics. There are a number of new developments in RNA silencing. Among the notable advances in the field is the progress in understanding of RNA silencing mechanisms and their regulations. The growing number of knock-out mammalian models is also positive and include conditional knock-outs of many RNA silencing components and specific knock-outs of miRNA clusters. Another remarkable trend is the combination of bioinformatic analysis with traditional approaches of molecular biology, which leads to better understanding of miRNA function in complex systems. Finally, the progress in oligonucleotide chemistry generates an unprecedented number of tools for sequence-specific gene silencing and for studies of RNA silencing.

Acknowledgements

The authors thank Javier Martinez and his team for organizing, for the third time, an excellent meeting, and look forward to an encore. We are grateful to the speakers for their highly informative talks and willingness to discuss unpublished data, some of which could not be featured in the report due to pending publications. We are also thankful to Donal O'Carroll for his amusing comment, which provided a title for this report. We also thank Donald Rumsfeld for inspiring Donal. Last, but not the least, we greatly appreciate support from sponsors who made this meeting possible.

C.K. is funded by Cancer Research UK and SNF, P.S. by the EMBO SDIG program and the Purkynje fellowship.

References

1. Gy I, Gasciolli V, Lauressergues D, Morel JB, Gombert J, Proux F, Proux C, Vaucheret H, Mallory AC. Arabidopsis FIERY1, XRN2 and XRN3 are endogenous RNA silencing suppressors. Plant Cell 2007; 19:3451-61.
2. Ramachandran V, Chen X. Degradation of microRNAs by a Family of Exoribonucleases in Arabidopsis. Science 2008; 321:1490-2.
3. Majoros WH, Ohler U. Spatial preferences of microRNA targets in 3' untranslated regions. BMC Genomics 2007; 8:152.
4. Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. Science 2008; 320:1643-7.
5. Selbach M, Schwanhausser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N. Widespread changes in protein synthesis induced by microRNAs. Nature 2008; 455:58-63.
6. O'Carroll D, Mecklenbrauker I, Das PP, Santana A, Koenig U, Enright AJ, Miska EA, Tarakhovsky A. A Slicer-independent role for Argonaute 2 in hematopoiesis and the microRNA pathway. Genes Dev 2007; 21:1999-2004.
7. Blenkiron C, Goldstein LD, Thorne NP, Spiteri I, Chin SF, Dunning MJ, Barbosa-Morais NL, Teschendorff AE, Green AR, Ellis IO, Tavare S, Caldas C, Miska EA. MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype. Genome Biol 2007; 8:214.
8. Newman MA, Thomson JM, Hammond SM. Lin-28 interaction with the let-7 precursor loop mediates regulated microRNA processing. RNA 2008; 14:1539-49.
9. Bussing I, Slack FJ, Grosshans H. let-7 microRNAs in development, stem cells and cancer. Trends Mol Med 2008; 14:400-9.
10. Giraldez AJ, Mishima Y, Rihel J, Grocock RJ, Van Dongen S, Inoue K, Enright AJ, Schier AF. Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. Science 2006; 312:75-9.
11. Tang F, Kaneda M, O'Carroll D, Hajkova P, Barton SC, Sun YA, Lee C, Tarakhovsky A, Lao K, Surani MA. Maternal microRNAs are essential for mouse zygotic development. Genes Dev 2007; 21:644-8.
12. Houbaviy HB, Murray MF, Sharp PA. Embryonic stem cell-specific MicroRNAs. Dev Cell 2003; 5:351-8.
13. Suh MR, Lee Y, Kim JY, Kim SK, Moon SH, Lee JY, Cha KY, Chung HM, Yoon HS, Moon SY, Kim VN, Kim KS. Human embryonic stem cells express a unique set of microRNAs. Dev Biol 2004; 270:488-98.
14. Mendell JT. miRiad roles for the miR-17-92 cluster in development and disease. Cell 2008; 133:217-22.

15. Ventura A, Young AG, Winslow MM, Lintault L, Meissner A, Erkeland SJ, Newman J, Bronson RT, Crowley D, Stone JR, Jaenisch R, Sharp PA, Jacks T. Targeted deletion reveals essential and overlapping functions of the miR-17 through 92 family of miRNA clusters. Cell 2008; 132:875-86.
16. Sinkkonen L, Hugenschmidt T, Berninger P, Gaidatzis D, Mohn F, Artus-Revel CG, Zavolan M, Svoboda P, Filipowicz W. MicroRNAs control de novo DNA methylation through regulation of transcriptional repressors in mouse embryonic stem cells. Nat Struct Mol Biol 2008; 15:259-67.
17. Choi WY, Giraldez AJ, Schier AF. Target protectors reveal dampening and balancing of Nodal agonist and antagonist by miR-430. Science 2007; 318:271-4.
18. Buhler M, Haas W, Gygi SP, Moazed D. RNAi-dependent and -independent RNA turnover mechanisms contribute to heterochromatic gene silencing. Cell 2007; 129:707-21.
19. Buhler M, Spies N, Bartel DP, Moazed D. TRAMP-mediated RNA surveillance prevents spurious entry of RNAs into the *Schizosaccharomyces pombe* siRNA pathway. Nat Struct Mol Biol 2008.
20. Houwing S, Kamminga LM, Berezikov E, Cronembold D, Girard A, van den Elst H, Filippov DV, Blaser H, Raz E, Moens CB, Plasterk RH, Hannon GJ, Draper BW, Ketting RF. A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. Cell 2007; 129:69-82.
21. Murchison EP, Stein P, Xuan Z, Pan H, Zhang MQ, Schultz RM, Hannon GJ. Critical roles for Dicer in the female germline. Genes Dev 2007; 21:682-93.
22. Aravin AA, Sachidanandam R, Girard A, Fejes-Toth K, Hannon GJ. Developmentally regulated piRNA clusters implicate MILI in transposon control. Science 2007; 316:744-7.
23. Tam OH, Aravin AA, Stein P, Girard A, Murchison EP, Cheloufi S, Hodges E, Anger M, Sachidanandam R, Schultz RM, Hannon GJ. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. Nature 2008; 453:534-8.
24. Sharma CM, Darfeuille F, Plantinga TH, Vogel J. A small RNA regulates multiple ABC transporter mRNAs by targeting C/A-rich elements inside and upstream of ribosome-binding sites. Genes Dev 2007; 21:2804-17.
25. Papenfort K, Pfeiffer V, Mika F, Lucchini S, Hinton JC, Vogel J. SigmaE-dependent small RNAs of Salmonella respond to membrane stress by accelerating global omp mRNA decay. Mol Microbiol 2006; 62:1674-88.
26. Tafer H, Ameres SL, Obernosterer G, Gebeshuber CA, Schroeder R, Martinez J, Hofacker IL. The impact of target site accessibility on the design of effective siRNAs. Nat Biotechnol 2008; 26:578-83.
27. Obernosterer G, Martinez J, Alenius M. Locked nucleic acid-based in situ detection of microRNAs in mouse tissue sections. Nat Protoc 2007; 2:1508-14.

Riboswitches: A Common RNA Regulatory Element

By: Andrea L. Edwards, B.S. & Robert T. Batey, Ph.D. (*Dept. of Chemistry and Biochemistry, University of Colorado*)

Citation: Edwards, A. L. & Batey, R. T. (2010) Riboswitches: A Common RNA Regulatory Element. *Nature Education* 3(9):9

RNA is able to control its own transcription, given the right trigger. See how the RNA-based riboswitch twists into a tangle in the presence of the right ligand.

Every living organism must be able to sense environmental stimuli and convert these input signals into appropriate cellular responses. Most of these responses are mediated by transcription factors that bind DNA and coordinate the activity of RNA polymerase or of proteins that elicit allosteric effects on their regulatory targets. In the early 1970s, researchers began to recognize that regions of mRNA transcripts have a regulatory role in the expression of downstream gene products (Jackson & Yanofsky 1973). By the early 1990s, several new regulatory mechanisms had been discovered that center on the action of RNA. (Arnaud *et al.* 1996; Aymerich & Steinmetz 1992; Houman *et al.* 1990; Lu *et al.* 1996; Oda *et al.* 2000; Wray & Fisher 1994). One classic example of regulation by RNA (often referred to as riboregulation) was discovered by Charles Yanofsky, who described the regulation of tryptophan biosynthesis at the mRNA level through the coupling of translation and transcription (Babtizke & Yanofsky 1993; Otridge & Gollnick 1993; Shimotsu *et al.* 1986; Yanofsky *et al.* 1996). Since then, many diverse RNA-based regulatory mechanisms have been discovered, including one that regulates interference and epigenetic regulation by long, noncoding RNA in eukaryotes (Costa 2007; Mattick 2001).

Riboregulation in Bacteria

Genetic regulation by RNA is widespread in bacteria. One common form of riboregulation in bacteria is the use of ribonucleic acid sequences encoded within mRNA that directly affect the expression of genes encoded in the full transcript (called cis-acting elements because they act on the same molecule they're coded in). These regulatory elements are known as riboswitches and are defined as mRNA elements that bind metabolites or metal ions as ligands and regulate mRNA expression by forming alternative structures in response to this ligand binding (Figure 1; Nudler & Mironov 2004; Tucker & Breaker 2005; Winkler 2005). To date, scientists have discovered that a variety of ligands are sensed by riboswitches; the group includes magnesium ions, nucleic acid precursors,

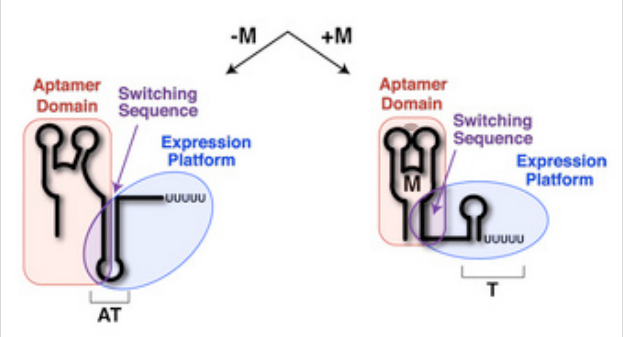


Figure 1: Riboswitch domains

A riboswitch can adopt different secondary structures to effect gene regulation depending on whether ligand is bound. This schematic is an example of a riboswitch that controls transcription. When metabolite is not bound (-M), the expression platform incorporates the switching sequence into an antiterminator stem-loop (AT) and transcription proceeds through the coding

enzyme cofactors, and amino acid residues (Table 1; Hammann & Westhof 2007).

Where are riboswitches located?

Riboswitches are most often located in the 5' untranslated region (5' UTR; a stretch of RNA that precedes the translation start site) of bacterial mRNA. There they regulate the occlusion of signals for transcription attenuation or translation initiation (Figure 2). However, not all riboswitches are at the 5' UTR; scientists have discovered that, in some eukaryotic mRNA, the thiamine pyrophosphate (TPP) riboswitch regulates splicing at the 3' end (Wachter *et al.* 2007).

region of the mRNA. When metabolite binds (+M), the switching sequence is incorporated into the aptamer domain, and the expression platform folds into a terminator stem–loop (T), causing transcription to abort. aptamer domain (red), switching sequence (purple), and expression platform (blue).

Table 1: Representative riboswitch classes

A broad overview of known riboswitches, including the classes that are the best understood to date.

Ligand type	Ligand name	Riboswitch family/class	Discovery (year)	Structure (year)
Enzyme cofactor	*S*-adenosylmethionine	SAM/SAM-I	2003 (Winkler *et al.* 2003)	2006 (Montange & Batey 2006)
	S-adenosylmethionine	SAM/SAM-II	2005 (Corbino *et al.* 2005)	2008 (Gilbert *et al.* 2008)
	S-adenosylmethionine	SAM/SAM-III	2006 (Fuchs *et al.* 2006)	2008 (Lu *et al.* 2008)
	Thiamine pyrophosphate	TPP	2001 (Miranda-Rios *et al.* 2001)	2006 (Serganov *et al.* 2006; Thore *et al.* 2006)
Nucleotide **precursor**	Guanine	Purine/G	2003 (Mandal *et al.* 2003)	2004 (Batey *et al.* 2004; Serganov *et al.* 2004)
	Adenine	Purine/A	2003 (Mandal & Breaker 2004)	2004 (Serganov *et al.* 2004)
	2'-Deoxyguanosine	Purine/dG	2007 (Kim *et al.* 2007)	2009 (Edwards & Batey 2009)
Amino acid	Lysine	Lysine	2003 (Rodionov *et al.* 2003; Sudarsan *et al.* 2003)	2008 (Garst *et al.* 2008; Serganov *et al.* 2008)
Metal ion	Magnesium	Mg^{2+}/ykoK	Identified 2004 (Barrick *et al.* 2004); validated 2007 (Dann *et al.* 2007)	2007 (Dann *et al.* 2007)

What Is the Structure of a Riboswitch?

The function of riboswitches is tied to the ability of RNA to form a diversity of structures. The most basic of these is the double-stranded helix, similar to that found in DNA. However, since most RNAs, unlike DNA, do not need to maintain perfect Watson-Crick base pairing, they can form other types of structures. For example, a single strand of RNA can fold back on itself to form a hairpin, which is composed of a helix capped by a loop. These elements of structure are called secondary structure. In larger RNAs, the helices and hairpins pack together into a specific pattern, referred to as the tertiary structure.

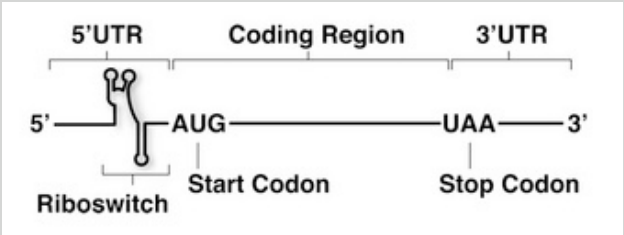


Figure 2: mRNA anatomy

A typical bacterial mRNA transcript controlled by a riboregulatory element such as a riboswitch is composed of three sections: the 5' untranslated region (5' UTR), the protein-coding region beginning with the start codon (AUG) and ending with a stop codon (UAA), and the 3' untranslated region (3' UTR).

Riboswitches are composed of two domains: the aptamer domain and the expression platform (Figure 1; Tucker & Breaker 2005). The aptamer domain acts as a receptor that specifically binds a ligand (Figure 1, red). The expression platform acts directly on gene expression through its ability to toggle between two different secondary structures in response to ligand binding (Figure 1, blue). Common to both domains is something called the switching sequence, and its placement in the aptamer domain or the expression platform ultimately dictates the expression outcome of the mRNA (Figure 1, purple). Specifically, if metabolite binding to the riboswitch stabilizes incorporation of the switching sequence into the aptamer domain, the expression platform must fold into a specific structure. Riboswitches that control transcriptional repression have a switching sequence that directs formation of a Rho-independent transcriptional terminator, a short stem-loop structure (followed by six or more uridine residues) that signals RNA polymerase to abort transcription (Nudler & Gottesman 2002; Hammann & Westhof 2007). Other riboswitches that regulate translational initiation utilize a switching sequence that can expose or occlude a ribosome-binding site (called the Shine-Dalgarno sequence; Hammann & Westhof 2007).

How Are Riboswitches Categorized?

Riboswitches are organized into families and classes according to two features: the type of ligand they bind, and their secondary structure (the arrangement of Watson-Crick paired helices; Hammann & Westhof 2007; Montange & Batey 2008). A family of riboswitches is typically a group of RNAs related by the ligands they recognize. For example, the SAM riboswitch family recognizes the compound *S*-adenosylmethionine (SAM). Within a family, there may be distinct classes of riboswitches, each class distinguished by a common sequence pattern that usually defines the ligand-binding pocket, as well as features required for folding the RNA into a three-dimensional shape. The SAM riboswitch family contains at least five known classes (Wang & Breaker 2008). These classes are distinguished from one another by their architectural features. For example, the SAM-I class forms a four-way helical junction, SAM-II forms a classic (H-type) pseudoknot, and SAM-III is

defined by a three-way junction (Figure 3; Montange & Batey 2006; Gilbert *et al.* 2008; Lu *et al.* 2008).

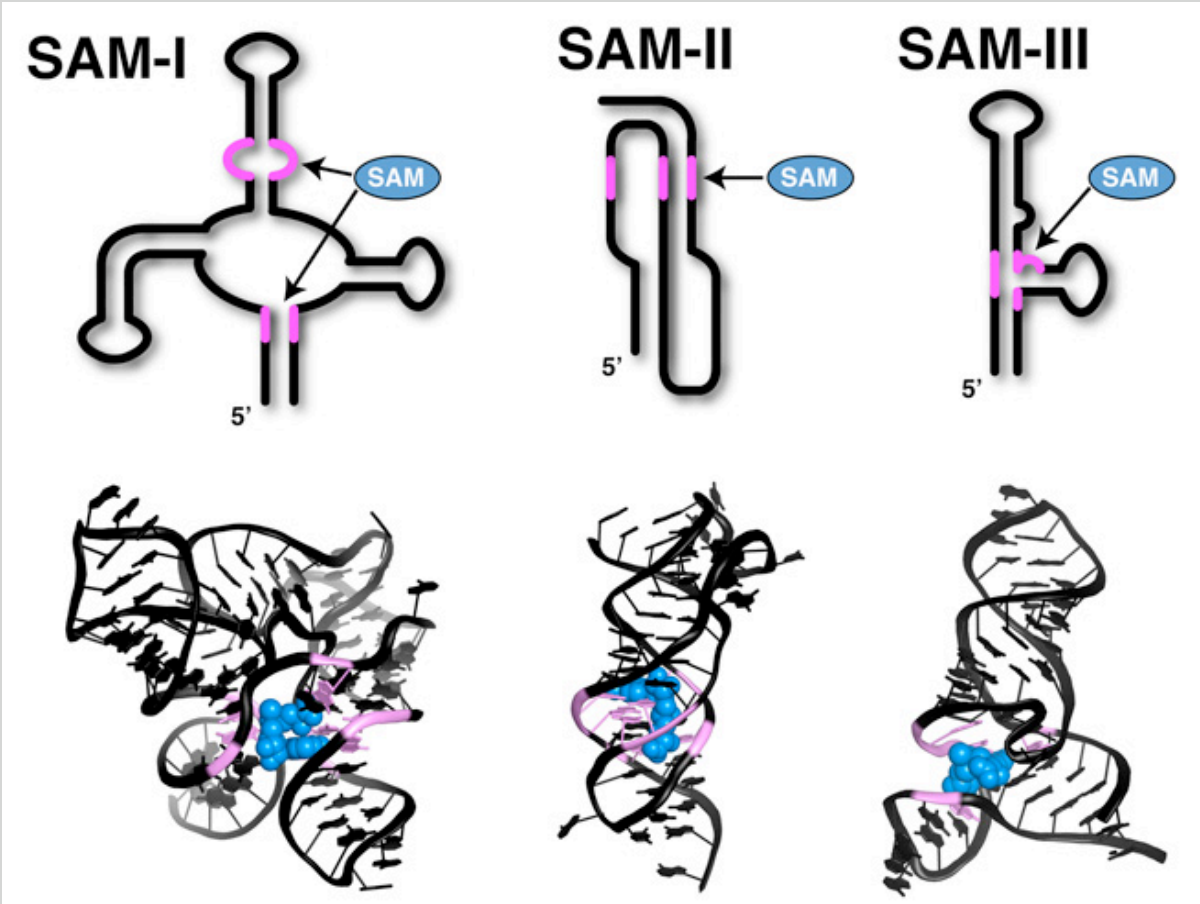


Figure 3: The SAM riboswitch family

The SAM-I, SAM-II, and SAM-III riboswitch classes are members of the SAM riboswitch family. The aptamer domain from each class has unique secondary (top row) and tertiary (bottom row) structural features. The ligand-binding pocket of each aptamer domain is shown in pink; the ligand (SAM) is shown in blue.

Purine Riboswitch Family: Global Structure

An exception to the classification scheme described above is the purine riboswitch family, a group of RNAs that actually share a common secondary structure but can recognize multiple distinct ligands (Kim & Breaker 2008).

The purine riboswitch family includes the adenine, guanine, and 2'-deoxyguanosine classes. Because its members recognize multiple ligands, this riboswitch family serves as a model for understanding the mechanisms of ligand recognition (Kim & Breaker 2008). The global architecture of the RNA in a purine riboswitch is defined by the organization of the three conserved helices that make up the secondary structure (Figure 4A). Two of the RNA helices form a coaxial stack, meaning that one helix sits on top of the other, and they are collinear. This pairing is the basis for their names, P1 and P3; P is an abbreviation for "paired" (Figure 4B). The third helix (P2) is adjacent to P3, and the terminal loops of P2 and P3 together form a tertiary structure called a loop-loop. (Because they form loops, P2 is also called L2, and P3 is also called L3 — since L is an abbreviation for "loop"). A complex

set of interacting helices and loop formations defines the overall three-dimensional fold of the purine riboswitch aptamer domain where ligands bind (Batey *et al.* 2004; Serganov *et al.* 2004).

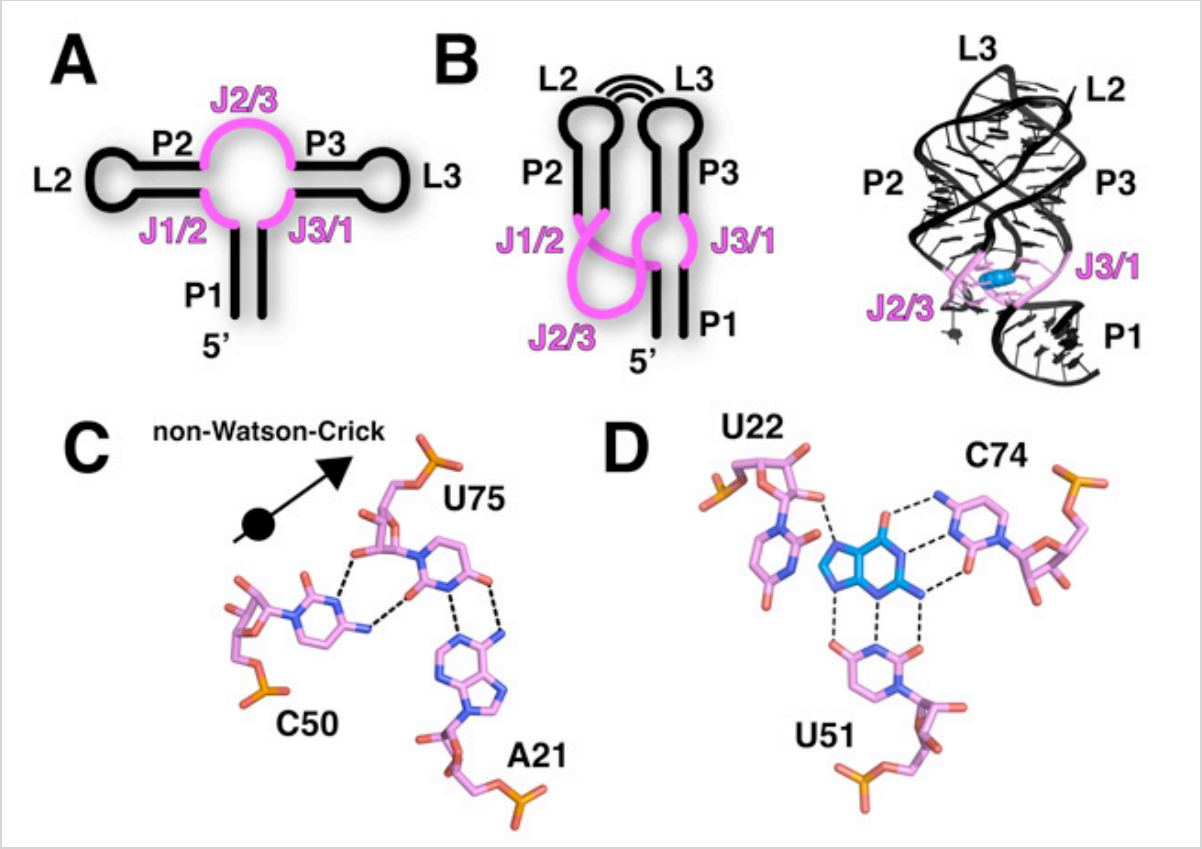


Figure 4: The purine riboswitch

The purine riboswitch family encompasses the guanine, adenine, and 2'-deoxyguanosine riboswitch classes. In all figures here, the binding pocket is highlighted in pink and the ligand, guanine, is shown in blue. A. The secondary structure of the purine riboswitch is shown with the paired regions (P1, P2, and P3), joining regions (J1/2, J2/3, and J3/1), and looped regions (L2 and L3) labeled. B. The left image is a schematic diagram of the secondary structure that emphasizes the tertiary interaction between L2 and L3, as well as the coaxial stack between P1 and P3. The right image shows the crystal structure of the ligand-bound guanine riboswitch. C. RNA commonly uses non-Watson-Crick interactions to stabilize tertiary architecture. The base triple among U75, A21, and C50 is shown. U75 and A21 form a canonical Watson-Crick pair, while the Watson-Crick face of C50 interacts with the sugar face of U75, forming a non-Watson-Crick interaction. D. Guanine specificity by the guanine riboswitch is derived from a Watson-Crick interaction between the ligand and the pyrimidine at position 74. Conserved uridine residues at positions 22 and 51 recognize all other functional groups on the ligand.

Purine Riboswitch Family: Ligand Binding in the Aptamer Domain

The three-way helical junction where P1, P2, and P3 meet is the ligand-binding pocket of a purine riboswitch. This region of the RNA is defined by a series of noncanonical base interactions (i.e., interactions that are not exclusively involved in Watson-Crick pairing). For example, in the P1 helix proximal to the ligand-binding site, a base triple interaction is observed in most purine riboswitches (Figure 4C). This triple, as well as most other unusual base triples, is typically composed of a Watson-Crick pair (A21-U75) interacting with a third

base (C50). At the center of the junction, a pyrimidine (Y) at position 74 forms a Watson-Crick pairing interaction with the ligand, which is further surrounded by other conserved residues (Figure 4D). The identity of this pyrimidine residue (cytosine or uridine) is the basis for specificity between the guanine and adenine classes (Noeske *et al.* 2005; Gilbert *et al.* 2006).

One of the most notable features of the ligand-bound aptamer domain is the way it nearly completely encapsulates the ligand, with the surrounding aqueous environment making the ligand almost entirely inaccessible (Batey *et al.* 2004; Serganov *et al.* 2004). Scientists have observed this encapsulation in other riboswitches, suggesting that this sequestering of the ligand is an important feature for the riboswitch mechanism of gene regulation (Edwards *et al.* 2007). The observation of this commonality also implies that ligand binding is often accompanied by a local conformational change in the aptamer domain, such that one element closes down on the binding pocket like a lid. What mechanism controls this lid closing? Scientists have used a combination of chemical probing experiments, fluorescence-based measurements, and kinetic analyses to look for an answer. Here is what some have proposed: one strand of the three-way junction, J3/1 (J for "joining") acts as a preformed docking station for the incoming purine ligand by making Y74 available for Watson-Crick pairing. Once the ligand is paired with Y74, nucleotides comprising J2/3 then fold over the bound ligand and encapsulate it (Gilbert *et al.* 2006; Noeske *et al.* 2007; Stoddard *et al.* 2008).

Purine Riboswitch Family: Ligand Binding and Regulatory Mechanism

The coupling of ligand binding to a conformational change is central to the regulatory mechanism of the riboswitch. As J2/3 encapsulates the ligand, it also forms additional tertiary interactions with the 3' strand of the P1 helix (including the base triple shown in Figure 4D). This region of the P1 helix is the switching sequence (Figure 1, purple). These additional ligand-induced interactions with the P1 helix stabilize its incorporation into the aptamer domain. Why is this important? It prevents the helix from being utilized to form alternative structures in the expression platform. In this way, the expression platform is fated to form one of two structures that interface with the expression machinery: either RNA polymerase or the ribosome. Thus, the ligand-binding event and the downstream regulatory switch are coupled through a limited ligand-induced conformational change that involves the switching sequence. This concept, originally proposed after scientists figured out the structure of the first riboswitch, has proven to be a nearly universal feature of riboswitches (Batey *et al.* 2004).

Kinetic Control of Riboswitches: A Race with Transcription

The name "riboswitch" seems to imply that when the RNA element is synthesized, it can reversibly switch back and forth between an on state and an off state, depending on the concentration of the ligand. When the ability to repress (or activate) gene expression is dictated by its inherent affinity for the ligand, a riboswitch is said to be thermodynamically controlled. In contrast, functional studies of a number of riboswitches have revealed that they

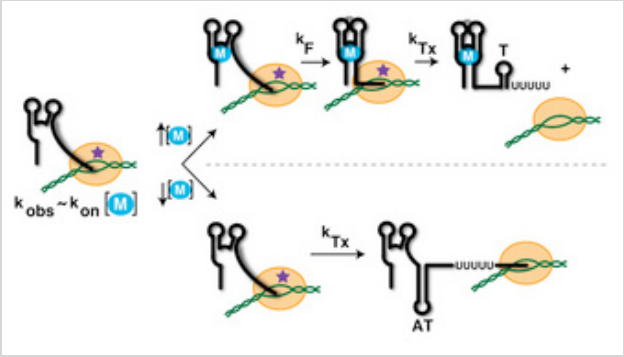


Figure 5: A model of kinetically

require a much higher ligand concentration to be activated, and alter the level of gene expression (Wickiser, Winkler, *et al.* 2005; Winkler 2005). Evidently, because riboswitches do not reach equilibrium with the ligand, the analogy is more like that of a fuse than a switch, so "ribo*fuse*" may be a better name.

Given these observations about ligand concentration, ligand–riboswitch binding, and riboswitch activation, scientists have proposed a more plausible model, stating that riboswitches are under kinetic control (Wickiser, Cheah, *et al.* 2005; Wickiser, Winkler, *et al.* 2005; Gilbert *et al.* 2006). Recall that a kinetically controlled event is one that depends largely or entirely on the rate-limiting step of a chemical process. For riboswitches, we must understand the relationship between the rates of transcription, RNA folding, and ligand binding. During transcription, the aptamer domain is always the first part of the RNA to be synthesized and fold into a shape capable of binding ligand. After completion of the aptamer domain, often a programmed pause site in the mRNA causes the polymerase to temporarily stall (Figure 5, star). This short pause (actually measured to be 2–10 seconds) gives the aptamer domain time to "interrogate" the cellular environment for the presence of ligand (Wickiser, Winkler, *et al.* 2005). If a sufficient concentration of ligand is present, then the ligand will occupy the binding pocket prior to resumption of transcription, and cause the mRNA to form a Rho-independent transcriptional terminator. Often a second pause site midway through the expression platform gives the RNA time, if ligand has not bound, to reconfigure the RNA secondary structure. In this case, the switching sequence will be used to form the antiterminator (Figure 5, AT), a stem-loop structure that disallows formation of the terminator, resulting in full transcription of the mRNA and allowing for its expression into protein.

controlled riboswitches
This cartoon shows a kinetically controlled riboswitch that regulates mRNA transcription. RNA polymerase (orange circle) transcribes the aptamer domain of a riboswitch (black) prior to transcription of the expression platform. A pause site (purple star) causes RNA polymerase to stall on the DNA template (green). The observed rate of ligand binding to the aptamer domain (kobs) is approximately equal to the rate constant associated with ligand binding (kon) multiplied by the ligand concentration ([M]). The rate of RNA folding (kF) in response to ligand binding must occur before RNA polymerase can fully transcribe the expression platform (kTx). The antiterminator and terminator stem-loops are labeled AT and T, respectively.

The kinetic control component of this process is the balance of rates of several processes: (1) speed of riboswitch transcription, including time spent residing at any potential pause sites; (2) ligand binding; and (3) a possible secondary structural rearrangement. Because several studies have shown that the rate of ligand binding to the aptamer domain is slow, the cellular concentration of ligand must be far greater than the dissociation constant (K_D) in order for binding to be sufficiently rapid to beat the kinetics of transcription (Wickiser, Cheah, *et al.* 2005; Wickiser, Winkler, *et al.* 2005; Gilbert *et al.* 2006).

How Riboswitches Can Affect Humans

Since these fascinating riboswitches are mechanisms specific to bacteria, it may be difficult determine how relevant they are to humans and human health. However, their role in regulating transcription in bacteria makes them enticing targets for the development of novel antibiotics aimed at stopping bacterial pathogens from flourishing inside the people they infect. Because riboswitches control genes essential for bacterial survival, or genes that

control the ability of bacteria to succeed at infection, a drug designed to affect a riboswitch could be a powerful tool for shutting down pathogenic bacteria (Blount & Breaker 2006). In fact, many antimicrobial compounds affect RNA directly, and many commonly used antibiotics inhibit translation by targeting bacterial ribosomes through binding interactions with ribosomal RNA (Vicens & Westhof 2003; Sutcliffe 2005). In addition, some compounds bind to the lysine, TPP, and FMN riboswitch classes and slow bacterial cell growth (Sudarsan *et al.* 2005; Blount *et al.* 2007; Lee *et al.* 2009).

Optimism for Novel Riboswitch-Based Therapeutics

There are two main reasons to expect that drugs designed to inhibit riboswitches would have minimal side effects in humans. One is that riboswitches have not been identified in mammals, so they are not likely to act on mammalian mRNA. Another is that some riboswitches are known to bind their cognate ligand in fundamentally different ways than do mammalian proteins that recognize the same ligand (Montange & Batey 2006). This means there is reason to suspect they won't interfere with ligand binding in native mammalian systems. Scientists have made some advances toward this kind of riboswitch-based drug design by figuring out the details of riboswitch-ligand complex structure. These structural data provide the basis for designing drugs that interfere with the riboswitch complexes and affect their shape, the fundamental basis for riboswitch control of RNA transcription.

References and Recommended Reading

Arnaud, M. *et al. In vitro* reconstitution of transcriptional antitermination by the SacT and SacY proteins of *Bacillus subtilis*. *Journal of Biological Chemistry* **271**, 18966–18972 (1996)

Aymerich, S. & Steinmetz, M. Specificity determinants and structural features in the RNA target of the bacterial antiterminator proteins of the BglG/SacY family. *PNAS* **89**, 10410–10414 (1992)

Babitzke, P. & Yanofsky, C. *PNAS* **90**, 133–137 (1993)

Batey, R. T., Gilbert, S. D. & Montange, R. K. Structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine. *Nature* **432**, 411–415 (2004)

Blount, K. F. & Breaker, R. R. Riboswitches as antibacterial drug targets. *Nature Biotechnology* **24**, 1558–1564 (2006)

Blount, K. F. *et al.* Antibacterial lysine analogs that target lysine riboswitches. *Nature Chemical Biology* **3**, 44–49 (2007)

Costa, F. F. Non-coding RNAs: Lost in translation? *Gene* **386**, 1–10 (2007)

Dann, C. E. *et al.* Structure and mechanism of a metal-sensing regulatory RNA. *Cell* **130**, 878–892 (2007)

Edwards, T. E., Klein, D. J. & Ferre-D'Amare, A. R. Riboswitches: Small-molecule recognition by gene regulatory RNAs. *Current Opinion in Structural Biology* **17**, 273–279 (2007)

Gilbert, S. D. *et al.* Structure of the SAM-II riboswitch bound to *S* -adenosylmethionine. *Nature Structural and Molecular Biology* **15**, 177–182 (2008)

Gilbert, S. D. *et al.* Thermodynamic and kinetic characterization of ligand binding to the purine riboswitch aptamer domain.

Journal of Molecular Biology **359**, 754–768 (2006)

Hammann, C. & Westhof, E. Searching genomes for ribozymes and riboswitches. *Genome Biology* **8**, 210 (2007)

Houman, F., Diaz-Torres, M. R. & Wright, A. Transcriptional antitermination in the *bgl* operon of E. coli is modulated by a specific RNA binding protein. *Cell* **62**, 1153–1163 (1990)

Jackson, E. N. & Yanofsky, C. Thr region between the operator and first structural gene of the tryptophan operon of may have a regulatory function. *Journal of Molecular Biology* **76**, 89–101 (1973)

Kim, J. N. & Breaker, R. R. Purine sensing by riboswitches. *Biology of the Cell* **100**, 1–11 (2008)

Lee, E. R., Blount, K. F. & Breaker, R. R. Roseoflavin is a natural antibacterial compound that binds to FMN riboswitches and regulates gene expression. *RNA Biology* **6**, 187–194 (2009)

Lu, C. *et al.* Crystal structures of the SAM-III/S(MK) riboswitch reveal the SAM-dependent translation inhibition mechanism. *Nature Structural and Molecular Biology* **15**, 1076–1083 (2008)

Lu, Y., Turner, R. J. & Switzer, R. L. Function of RNA secondary structures in transcriptional attenuation of the operon. *PNAS* **93**, 14462–14467 (1996)

Mattick, J. S. Non-coding RNAs: The architects of eukaryotic complexity. *EMBO Reports* **2**, 986–991 (2001)

Montange, R. K. & Batey, R. T. Riboswitches: Emerging themes in RNA structure and function. *Annual Reviews of Biophysics* **37**, 117–133 (2008)

Montange, R. K. & Batey, R. T. Structure of the *S* -adenosylmethionine riboswitch regulatory mRNA element. *Nature* **441**, 1172–1175 (2006)

Noeske, J. *et al.* An intermolecular base triple as the basis of ligand specificity and affinity in the guanine- and adenine-sensing riboswitch RNAs. *PNAS* **102**, 1372–1377 (2005)

Noeske, J. *et al.* Interplay of "induced fit" and preorganization in the ligand induced folding of the aptamer domain of the guanine binding riboswitch. *Nucleic Acids Research* **35**, 572–583 (2007)

Nudler, E. & Gottesman, M. E. Transcription termination and anti-termination in *E. coli*. *Genes to Cells* **7**, 755–768 (2002)

Nudler, E. & Mironov, A. S. The riboswitch control of bacterial metabolism. *Trends in Biochemical Sciences* **29**, 11–17 (2004)

Oda, M. *et al.* operon and histidine-dependent binding of HutP to the transcript containing the regulatory sequences. *Molecular Microbiology* **35**, 1244–1254 (2000)

Otridge, J. & Gollnick, P. MtrB from RNA in a tryptophan-dependent manner. *PNAS* **90**, 128–132 (1993)

Serganov, A. *et al.* Structural basis for discriminative regulation of gene expression by adenine- and guanine-sensing mRNAs. *Chemistry and Biology* **11**, 1729–1741 (2004)

Shimotsu, H. *et al.* Novel form of transcription attenuation regulates expression the tryptophan operon. *Journal of Bacteriology* **166**, 461–471 (1986)

Stoddard, C. D., Gilbert, S. D. & Batey, R. T. Ligand-dependent folding of the three-way junction in the purine riboswitch. *RNA* **14**, 675–684 (2008)

Sudarsan, N. *et al.* Thiamine pyrophosphate riboswitches are targets for the antimicrobial compound pyrithiamine.

Chemistry and Biology **12**, 1325–1335 (2005)

Sutcliffe, J. A. Improving on nature: Antibiotics that target the ribosome. *Current Opinion in Microbiology* **8**, 534–542 (2005)

Tucker, B. J. & Breaker, R. R. Riboswitches as versatile gene control elements. *Current Opinion in Structural Biology* **15**, 342–348 (2005)

Vicens, Q. & Westhof, E. RNA as a drug target: The case of aminoglycosides. *Chembiochem* **4**, 1018–1023 (2003)

Wachter, A. *et al.* Riboswitch control of gene expression in plants by splicing and alternative 3' end processing of mRNAs. *Plant Cell* **19**, 3437–3450 (2007)

Wang, J. X. & Breaker, R. R. Riboswitches that sense . *Biochemistry and Cell Biology* **86**, 157–168 (2008)

et al. The kinetics of ligand binding by an adenine-sensing riboswitch. *Biochemistry* **44**, 13404–13414 (2005)

Wickiser, J. K., Winkler, W. C. *et al.* The speed of RNA transcription and metabolite binding kinetics operate an FMN riboswitch. *Molecular Cell* **18**, 49–60 (2005)

Winkler, W. C. Riboswitches and the role of noncoding RNAs in bacterial metabolic control. *Current Opinion in Chemical Biology* **9**, 594–602 (2005)

Wray, L. V., Jr. & Fisher, S. H. Analysis of Bacillus subtilis hut operon expression indicates that histidine-dependent induction is mediated primarily by transcriptional antitermination and that amino acid repression is mediated by two mechanisms: Regulation of transcription initiation and inhibition of histidine transport. *Journal of Bacteriology* **176**, 5466–5473 (1994)

Yanofsky, C., Konan, K. V. & Sarsero, J. P. Some novel transcription attenuation mechanisms used by bacteria. *Biochimie* **78**, 1017–1024 (1996)

Outline | **Keywords** | **Add Content to Group** FEEDBACK

Primer

RNA splicing

Andy Newman

In the 20 years since the discovery that many eukaryotic genes are interrupted by non-coding intervening sequences, or introns, biologists have learnt a tremendous amount about the nuclear machinery that removes these introns from precursor gene transcripts before they are translated into proteins. This nuclear machinery must recognize the introns in the mRNA transcripts and then excise them by means of mRNA splicing. Short conserved sequences at the ends of introns — splice sites — are crucial for intron recognition and for the accuracy of the splicing reactions. In the 1980s the development of biochemical systems for splicing mRNA precursors allowed the splicing machinery to be dissected and analysed.

Splicing occurs in spliceosomes, large particles which are built up stepwise on the mRNA precursor from smaller RNA–protein sub-assemblies called snRNPs (small nuclear ribonucleoprotein particles). Each of the snRNPs that makes up the spliceosome contains a small nuclear RNA (snRNA) and a common set of snRNP proteins, plus up to ten additional snRNP-specific proteins. A spliceosome is therefore composed of dozens of proteins in addition to the five spliceosomal snRNAs (the U1, U2, U4, U5 and U6 snRNAs), so that spliceosomes rival ribosomes in size and complexity.

Energy from the hydrolysis of ATP is required at several points during the formation and operation of the spliceosome. But although the spliceosome uses ATP as fuel, the actual chemical mechanism of mRNA splicing consists of two successive trans-esterification reactions, which do not themselves involve ATP hydrolysis (see Figure 1). Interestingly, a similar reaction scheme is used by a widespread class of highly structured so-called 'autocatalytic introns' — group II introns — which can self-splice without ATP hydrolysis or any trans-acting factors. This begs the questions: what are all those spliceosomal RNA and protein components doing; and why do spliceosomes require energy input from ATP? There is still much detail to unravel but the answers to these questions are beginning to emerge.

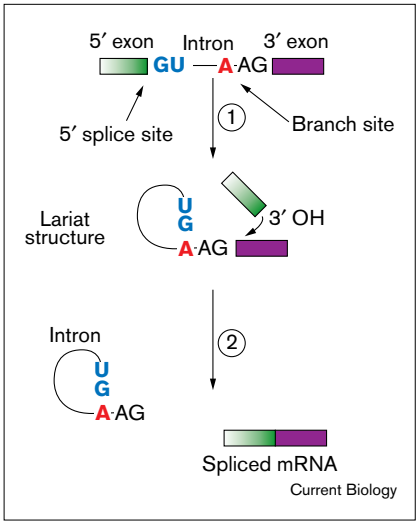
Molecular motors

The network of interactions between the snRNAs and the substrate is radically remodelled as the spliceosome is assembled and activated (see blue box). How are these rearrangements driven and orchestrated? Over the past few years biochemical purification of snRNPs and spliceosomes, and sequencing of the protein components, has led to the isolation of the genes or cDNAs that encode many of the splicing factors. Other spliceosome components have been identified by genetic approaches in yeast, and in many cases there are clear homologues in the yeast and mammalian systems. For some of these splicing factors the presence of specific sequence motifs provides clues to potential functions.

Several splicing factors have attracted particular attention because they include motifs present in ATP-powered RNA helicases, which are known to modify RNA structure in other multi-component systems such as the translation machinery. Recently, three of these helicase-like factors (Prp16 and Prp22 in yeast, and the 200 kDa U5 snRNP protein in mammals) have been shown to have ATP-dependent RNA unwinding activity. So far these factors have not been definitively linked to any particular spliceosomal rearrangement; substrate specificity may be determined by their local environment in the spliceosome. In any case the activity of these ATP-dependent helicases explains, at least in part, the requirement for ATP hydrolysis in spliceosomes.

Intriguingly, a recently identified spliceosomal protein that is a component of the U5 snRNP turns out to be highly homologous to EF-2, a component of the translation machinery, and like its ribosomal counterpart the spliceosomal protein is a GTPase. EF-2 is thought to couple GTP hydrolysis to the movement of peptidyl-tRNA during translation. The U5 snRNP GTPase might function in a similar way, perhaps to force some rearrangement or

Figure 1

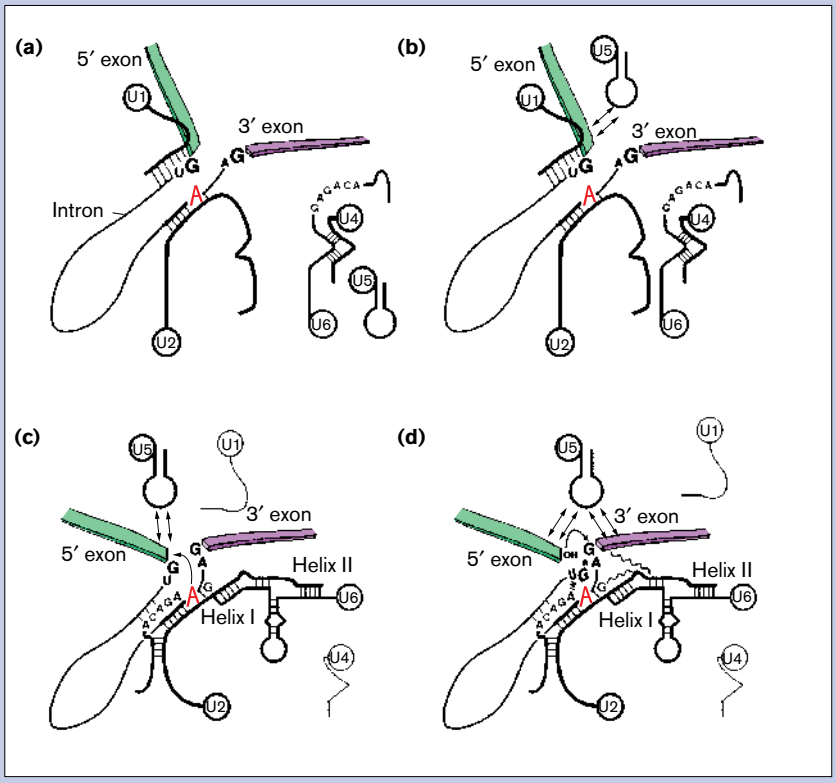


Splicing of mRNA precursors involves two successive trans-esterification reactions. In the first reaction (**1**) the 2′OH of a specific adenosine (red) at the branch site near the 3′ end of the intron attacks the 5′ splice site (blue). This reaction releases the 5′ exon (green; with a 3′ OH terminus) and leaves the 5′ end of the intron (blue) joined by a 2′–5′ phosphodiester bond to the the branch site adenosine (red); this intron–3′ exon intermediate is therefore in the form of a lariat. In the second reaction (**2**) the 3′ OH of the 5′ exon intermediate (green) attacks the 3′ splice site, producing the spliced mRNA and lariat-shaped intron products.

RNA networking

During mRNA splicing in the spliceosome, protein factors and short conserved RNA sequences in the spliceosomal snRNAs collaborate to identify the intron 5′ splice site and the branch site. The branch site is a specific adenosine residue near the 3′ end of the intron that provides the 2′ OH nucleophile for the first trans-esterification reaction. The 5′ splice site base-pairs with U1 snRNA early in spliceosome assembly and the branch site interacts with an invariant sequence in U2 snRNA. Base-pairing between U2 and the branch site bulges the branch site adenosine out of the helix (see Figure 2a). When U6 snRNA enters the spliceosome, it is extensively base-paired with U4 snRNA in the form of a U4–U6·U5 tri-snRNP particle. Before the first catalytic step the U4–U6 interaction is disrupted and an invariant motif in U6 supplants U1 snRNA at the 5′ splice site (see Figure 2c). A specific base-pairing interaction between U2 and U6 snRNAs (helix I; see Figure 2c) then juxtaposes the branch site adenosine and its 5′ splice site target. The U2–U6 sequences in the vicinity of helix I are some of the most conserved in the spliceosome and mutation of several of these nucleotides can block the first or second catalytic step of splicing, suggesting that these sequences perhaps form part of the spliceosome's active core. Meanwhile, U5 snRNP holds onto the 5′ exon intermediate, which has been cut free, and the spliceosome is reconfigured, in a way that is still poorly understood, to bring the 3′ splice site into the catalytic core and align the exons for the second catalytic step.

Figure 2



(a) Before the first trans-esterification, U1 snRNA and U2 snRNA base-pair to the 5′ splice site and intron branch site (red A), respectively. U4 and U6 snRNAs are base-paired together in a tri-snRNP particle with U5. **(b)** The loop sequence in U5 snRNA contacts the 5′ exon in the mRNA precursor. **(c)** The first trans-esterification. The U4–U6 interaction has been disrupted and U6 interacts with both U2 snRNA and the 5′ splice site, while U5 maintains contact with the 5′ exon. **(d)** The second trans-esterification. The network of interactions is reconfigured to juxtapose the 3′OH of the 5′ exon with the 3′ splice site, and U5 snRNA contacts both exons. (Adapted, with permission, from Nilsen TW in *RNA Structure and Function*, Cold Spring Harbor Press 1998.)

translocation of substrate sequences in the spliceosome.

Splice site recognition

It is axiomatic that mRNA splicing needs to be accurate and reliable; errors would produce mutated or truncated proteins, which could have catastrophic biological effects. In yeast the splice sites and branch site are highly conserved, and recognition of these signals might be sufficient to determine unambiguously the intron–exon boundaries. In most introns in higher eukaryotes, however, the equivalent sequences are typically quite degenerate; only the GU and AG dinucleotides at the intron termini are invariant. These sequences alone are clearly not sufficient to determine the intron boundaries but sequence analysis has identifed other sources of information in mRNA precursor sequences, such as different base composition, which might allow introns and exons to be distinguished. Also, there is compelling evidence for direct interactions between spliceosome components and proteins involved in 5′ cap formation and 3′ polyadenylation, so the signals which determine these other RNA processing steps might also influence splicing patterns.

A class of proteins called SR proteins, which contain RS domains (regions rich in arginine–serine dipeptides) together with RNA-binding domains, have critical roles in intron–exon definition in higher eukaryotes. They have been shown to recruit U1 snRNP to 5′ splice sites via specific protein–protein contacts with the U1 snRNP proteins. Other SR proteins form a network of RNA–protein and protein–protein interactions, which stabilises factors bound near the 5′ and 3′ splice sites. This bridging function plays a crucial part in early spliceosome formation and helps to define intron–exon boundaries. SR proteins also have fundamental roles in alternative splicing in higher

eukaryotes, typically by interaction with exonic splicing 'enhancer' sequences and by networking with factors bound to nearby splice sites to modulate splice site choice.

A new type of spliceosome

One of the most surprising and significant findings in recent years is the discovery of a new and distinct type of spliceosome in higher eukaryotes. Explosive growth in the amount of genomic sequence in databases has uncovered rare introns with non-consensus splice sites. These introns have characteristic, tightly conserved splice sites and branch site sequences that are quite distinct from those recognised by the conventional splicing machinery. They are processed by a novel spliceosome assembled from unusual, low-abundance snRNAs, which are functionally analogous to U1, U2, U4 and U6 snRNAs. Curiously, the novel spliceosomes contain U5 snRNA in common with conventional spliceosomes.

Splicing in these novel spliceosomes proceeds by the usual two-step trans-esterification pathway. The rare intron 5′ splice site and branch site sequences base-pair with U11 and U12 snRNAs (instead of the U1 and U2 snRNAs as they do in the conventional U2 spliceosomes) and novel U4-like and U6-like snRNAs substitute for conventional U4 and U6.

Interestingly, although the novel U4 and U6 snRNAs are not extensively similar in sequence to their conventional counterparts, they form a base-paired complex just like U4–U6 and almost certainly dissociate in the assembled U12 spliceosome. Furthermore, comparison of conventional U6 with its novel spliceosome counterpart shows that there are clear patches of sequence identity which correspond precisely to those nucleotides shown to be critical for U6 function in conventional spliceosomes.

In addition, the novel U6 snRNA can interact with U12 snRNA to form a base-paired structure similar to U2–U6 helix I. So, it looks as if the pattern of snRNA interactions in the U12 spliceosome is strikingly similar to that in the conventional U2 spliceosome, even though many of the RNA sequences are quite different. This implies that the two types of spliceosome function in similar ways and it strongly supports the current picture of snRNA interactions in U2 spliceosomes (see blue box). At present little is known about the identities of the protein components of U12 spliceosomes, but it will be interesting to compare these with their conventional U2 spliceosome counterparts.

Catalysis by RNA?

One of the major unanswered questions about spliceosome function concerns the nature of the catalytic core of the machinery. Are the reactions catalyzed by RNA or by protein, or perhaps both? There is a great deal of biochemical and mutational evidence implicating snRNA sequences — in particular, highly conserved regions of U2 and U6 in conventional spliceosomes — in the trans-esterification reactions, but no definitive proof that they contribute to the active site. In fact, it is not clear whether the spliceosome has one active site (which would, perhaps, be remodelled between the first and second reaction) or two sites, one for each of the chemical steps.

It has recently been shown that the spliceosome catalyses 5′ splice site cleavage via a divalent metal ion-dependent pathway, which would be consistent with RNA catalysis but by no means excludes other possibilities. Unraveling the contributions of the RNA and protein components of the machinery to the catalysis of splicing is clearly going to occupy molecular biologists for some time.

Accurate structural analysis of the spliceosome is still in its infancy but is going to be essential if splicing is to be understood in detail. The complexity and dynamic nature of the machinery will pose a major challenge for structural biologists in the next few years.

Key references

Nilsen TW: **RNA–RNA interactions in nuclear pre-mRNA splicing.** In *RNA Structure and Function.* New York: Cold Spring Harbor Press; 1998:279-307.

Sharp PA, Burge CB: **Classification of introns: U2-type or U12-type.** *Cell* 1997, **91:**875-879.

Staley JP, Guthrie C: **Mechanical devices of the spliceosome: motors, clocks, springs and things.** *Cell* 1998, **92:**315-326.

Tarn WY, Steitz JA: **Pre-mRNA splicing: the discovery of a new spliceosome doubles the challenge.** *Trends Biochem Sci* 1997, **22:**132-137.

Wang J, Manley JL: **Regulation of pre-mRNA splicing in metazoa.** *Curr Opin Gen & Dev* 1997, **7:**205-211.

Will CL, Luhrmann R: **Protein functions in pre-mRNA splicing.** *Curr Opin Cell Biol* 1997, **9:**320-328.

Address: Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK.

LETTERS

Relationship between nucleosome positioning and DNA methylation

Ramakrishna K. Chodavarapu[1]*, Suhua Feng[1,2]*, Yana V. Bernatavichute[1,3], Pao-Yang Chen[1], Hume Stroud[1], Yanchun Yu[4], Jonathan A. Hetzel[1], Frank Kuo[1], Jin Kim[1], Shawn J. Cokus[1], David Casero[1], Maria Bernal[5], Peter Huijser[6], Amander T. Clark[1,7], Ute Krämer[5]†, Sabeeha S. Merchant[3,8], Xiaoyu Zhang[4], Steven E. Jacobsen[1,2,3] & Matteo Pellegrini[1,3]

Nucleosomes compact and regulate access to DNA in the nucleus, and are composed of approximately 147 bases of DNA wrapped around a histone octamer[1,2]. Here we report a genome-wide nucleosome positioning analysis of *Arabidopsis thaliana* using massively parallel sequencing of mononucleosomes. By combining this data with profiles of DNA methylation at single base resolution, we identified 10-base periodicities in the DNA methylation status of nucleosome-bound DNA and found that nucleosomal DNA was more highly methylated than flanking DNA. These results indicate that nucleosome positioning influences DNA methylation patterning throughout the genome and that DNA methyltransferases preferentially target nucleosome-bound DNA. We also observed similar trends in human nucleosomal DNA, indicating that the relationships between nucleosomes and DNA methyltransferases are conserved. Finally, as has been observed in animals, nucleosomes were highly enriched on exons, and preferentially positioned at intron–exon and exon–intron boundaries. RNA polymerase II (Pol II) was also enriched on exons relative to introns, consistent with the hypothesis that nucleosome positioning regulates Pol II processivity. DNA methylation is also enriched on exons, consistent with the targeting of DNA methylation to nucleosomes, and suggesting a role for DNA methylation in exon definition.

To investigate the position of nucleosomes in *Arabidopsis thaliana*, we sequenced micrococcal nuclease (MNase)-digested nucleosomal DNA using an Illumina GAII sequencer to achieve a roughly 68-fold coverage of nucleosomes (see Supplementary Methods). The data are displayed in a modified UCSC genome browser (http://epigenomics.mcdb.ucla.edu/Nuc-Seq/) along with nucleosome density tracks that allow easy visualization of nucleosome positions throughout the genome (Fig. 1a).

To obtain a chromosomal view of nucleosome content, we plotted reads in bins of 100 kilobases tiling the chromosomes. As a control for biases in mapping and sequencing we also sequenced a library of randomly sheared *Arabidopsis* genomic DNA. We then normalized the nucleosome counts by the number of uniquely mapping genomic DNA counts within each bin along the chromosomes. Nucleosome content was relatively uniform throughout the euchromatic regions of chromosomes, but showed significant enrichment in pericentromeric heterochromatin regions, (Fig. 1b, Supplementary Fig. 1). To confirm these results, we performed chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-Seq) using an antibody against unmodified histone H3 (Fig. 1b and Supplementary Fig. 1). These observations indicate that nucleosomes are more densely packed in pericentromeric regions that are transcriptionally silent, are rich in transposons and contain heavily methylated DNA, than in the euchromatic arms[3].

By examining the relationship between reads that map to opposite strands of DNA, we observed a peak of reads on the reverse strand that occurred approximately 145–170 bases downstream of the reads on the forward strand (Supplementary Fig. 2a). This broad peak in the general vicinity of the accepted size of a nucleosome indicates that many nucleosomes are well positioned, leading to the repeated sequencing of both the forward and reverse complement of the corresponding nucleosome regions. Similarly, when we plotted the correlation between positive strand reads with other positive strand reads we saw a progressively decreasing correlation from the start of the nucleosome (Fig. 1c and Supplementary Fig. 2b), with smaller peaks spaced at 174 and 355 base pairs from the starting position of the reference reads indicating some preference for regular spacing of nucleosomes genome-wide. Assuming that nucleosomes are wrapped around 147 base pairs of DNA, the average length of the linkers of regularly spaced nucleosomes is about 30 base pairs. Consistent with the higher content of nucleosomes in pericentromeric heterochromatin regions, we found that the peaks at 174 and 355 were more pronounced in these regions than in the euchromatic arms (Fig. 1c), indicating that pericentromeric nucleosomes have a higher tendency to be in regularly spaced arrays. These results are consistent with earlier evidence for more regular spacing of nucleosomes in *Drosophila* heterochromatin[4].

Similar to findings in animal and fungal high-throughput nucleosome sequencing studies[5–8], we found 10-base periodicities in WW (W = A or T) dinucleotides, and SS (S = G or C) dinucleotides that were 5 bases out of phase with the WW dinucleotides (Fig. 1d and Supplementary Figs 3–5). WW dinucleotides are favoured at sites where the minor groove faces the histone core, whereas SS dinucleotides are favoured at sites where the minor groove faces away from the histone core[9,10]. The out of phase peaks in the frequencies of these dinucleotides leads to optimal bending of DNA, as A/T nucleotides cause a negative base roll and G/C nucleotides cause a positive base roll[9,10].

To study the relationship of DNA methylation with nucleosome positions, we used our single-nucleotide resolution whole-genome bisulphite-sequencing data[3]. *Arabidopsis* cytosines are methylated by

[1]Department of Molecular, Cell, and Developmental Biology, University of California Los Angeles, Los Angeles, California 90095, USA. [2]Howard Hughes Medical Institute, University of California Los Angeles, Los Angeles, California 90095, USA. [3]Molecular Biology Institute, University of California Los Angeles, Los Angeles, California 90095, USA. [4]Department of Plant Biology, University of Georgia, Athens, Georgia 30602, USA. [5]University of Heidelberg, BIOQUANT 23, R. 645, Im Neuenheimer Feld 267, D-69120 Heidelberg, Germany. [6]Department of Molecular Plant Genetics, Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany. [7]Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, University of California, Los Angeles, Los Angeles, California 90095, USA. [8]Department of Chemistry and Biochemistry, University of California, Los Angeles, Los Angeles, California 90095, USA. †Present address: Department of Plant Physiology, University of Bochum, Universitaetsstrasse 150, D-44801 Bochum, Germany.
*These authors contributed equally to this work.

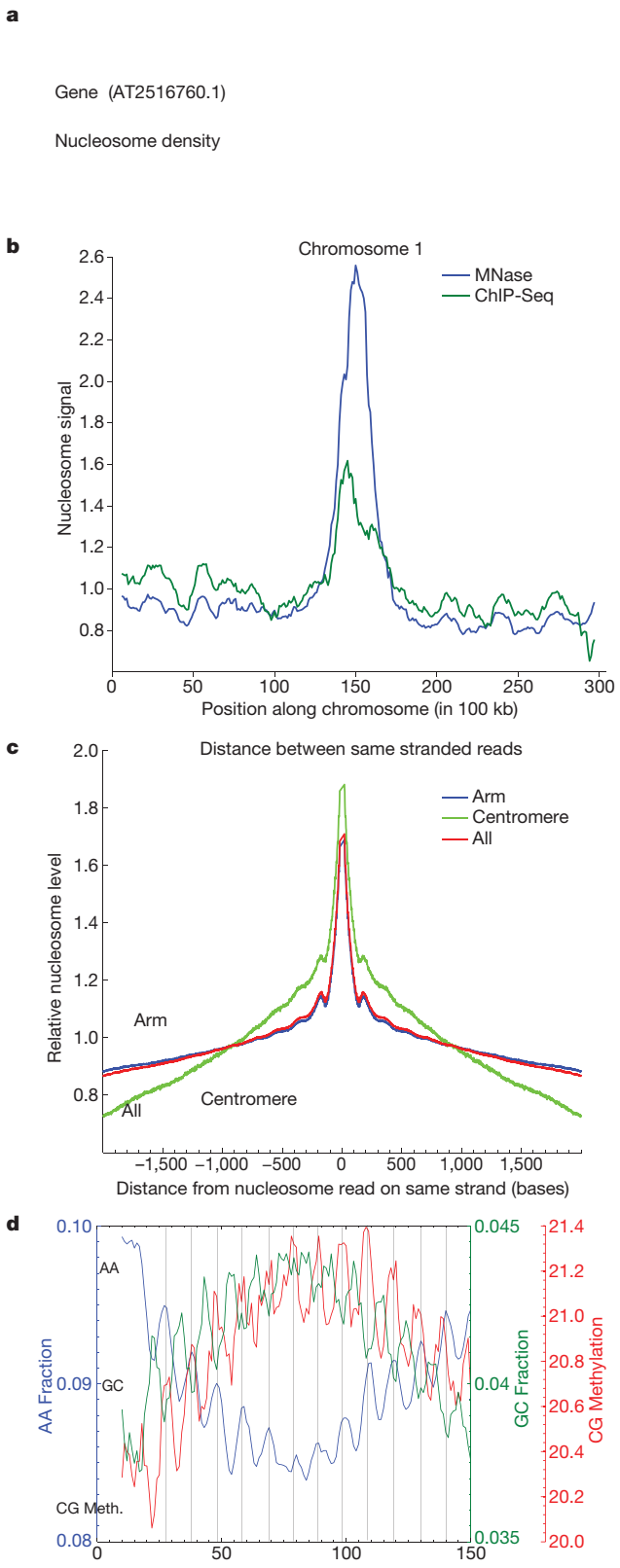


Figure 1 | Characterization of *A. thaliana* nucleosome data. **a**, Representative UCSC Browser screenshot of a gene showing calculated nucleosome densities. **b**, Chromosomal view of nucleosome read counts and ChIP-seq read counts in 100-kb tiles along chromosome 1 shows nucleosome enrichment in pericentromeric regions. **c**, Autocorrelation of positive stranded reads shows local peaks at positions 174 and 355 bases. **d**, AA and GC dinucleotide and CG methylation profiles show a 10-base periodicity over nucleosome-bound DNA, with DNA methylation profiles in phase with WW dinucleotide profiles, and out of phase with SS dinucleotide profiles.

one of three different DNA methyltransferases depending on their sequence context. CG sites are methylated by MET1, and CHG sites (where H is A, C or T) are methylated by CMT3. Finally, CHH sites are methylated by DRM2, a *de novo* methyltransferase that is targeted by small RNAs[11]. Notably, all three types of methylation showed a 10-base periodicity on nucleosomal DNA, which was in phase with the WW dinucleotides and out of phase with the SS dinucleotides (Figs 1d, 2a–c and Supplementary Fig. 6). These methylation preferences were not correlated with preferences for CG, CHG or CHH sequences at these locations (Supplementary Fig. 7). Because DNA methyltransferases access the major groove, this methylation would be on DNA that is on the outside of the nucleosome (minor groove facing the histones) and thus more accessible to the DNA methyltransferases. This in turn indicates that DNA might be in part methylated as the DNA is still bound to nucleosomes, leading to the observed 10-base pair periodicity. It has been proposed that chromatin remodelling enzymes are required for DNA methyltransferases to gain access to the DNA, and indeed DRD1, DDM1 and LSH1 are remodellers known to be important factors controlling DNA methylation[12]. However, our data indicate that nucleosomal DNA may also be a substrate for DNA methyltransferases *in vivo*, demonstrating a prominent role of the nucleosome in determining methylation patterning throughout the genome.

We previously reported a 10-nucleotide periodicity in CHH methylation data when performing autocorrelation analysis on the methylation pattern of the whole genome[3]. Our previous interpretation was that the structure of the DRM2 enzyme might be responsible for this pattern, because the orthologous Dnmt3 enzymes in mammals are known to act as heteromeric complexes in which two methyltransferase active sites have a spacing equivalent to roughly 10 nucleotides of DNA[13]. However, the current data in which we see 10-base pair periodicities for all types of methylation indicate a more general explanation: that nucleosomes are to some extent dictating access to the DNA and therefore setting the register of methylation for all DNA methyltransferases. Nucleosomal preferences could also partially explain the sequence preferences that we observed previously for CHG and CHH methylation[3]. Highly methylated cytosines tended to be followed immediately by A/T but not C, consistent with our finding that DNA methylation is out of phase with CC dinucleotides (Supplementary Fig. 6).

On a larger scale, we also observed that levels of all three types of DNA methylation were higher in nucleosome-spanning DNA than in flanking DNA, indicating that nucleosome-bound DNA is enriched for DNA methylation (Fig. 2 and Supplementary Fig. 8). This finding supports the view that nucleosomes are preferentially targeted by DNA methyltransferases. In the case of CMT3, it is predicted that this enzyme is recruited or activated by histone H3 lysine 9 dimethylation, because a knockout of histone H3 lysine 9 methyltransferase mimics a knock out of CMT3, and because the CMT3 chromodomain can bind to methylated histones[14]. However, our data show that all types of methylation are enriched on nucleosome-bound DNA, indicating that nucleosomes or histone modifications may assist in the recruitment of all of the *Arabidopsis* DNA methyltransferases.

To test whether the patterns of DNA methylation on nucleosomal DNA are also found in humans, we used previously published MNase nucleosome sequencing data[15] together with our single-nucleotide resolution bisulphite sequencing data on the human embryonic stem cell line HSF1 (see Supplementary Methods). We found that human nucleosome-bound DNA also showed a 10-base periodicity in its CG, CHG and CHH methylation status (Fig. 3). Furthermore, as in *Arabidopsis*, the overall level of methylation was higher on nucleosome-bound DNA than in flanking regions. This indicates that nucleosome architecture has a role in shaping DNA methylation patterning in the human genome, and is consistent with the recent finding of stable anchoring of Dnmt3 DNA methyltransferases to mammalian chromatin[16] and more generally with the conservation of DNA methyltransferase function in plants and animals[17]. We also

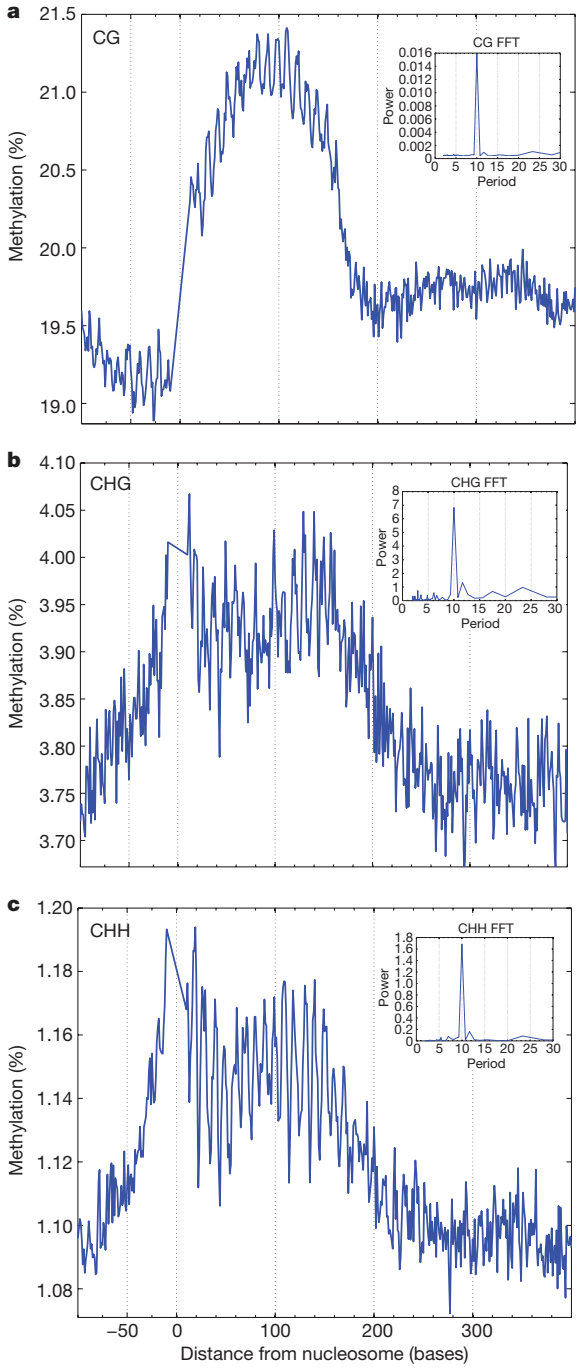


Figure 2 | DNA methylation profiles of nucleosome-bound DNA in *Arabidopsis*. The weighted average percent DNA methylation was calculated and plotted at each distance from nucleosome start sites (0). **a**–**c**, CG methylation (**a**), CHG methylation (**b**) and CHH methylation (**c**) each show a 10-base periodicity over nucleosome-bound DNA (1–147 bases). Fast Fourier transforms (FFTs) can be used to deconstruct inherent frequencies in a complex signal. The FFTs calculated over the region of the nucleosome (insets) demonstrate this periodicity in CG (**a**), CHG (**b**) and CHH (**c**) contexts.

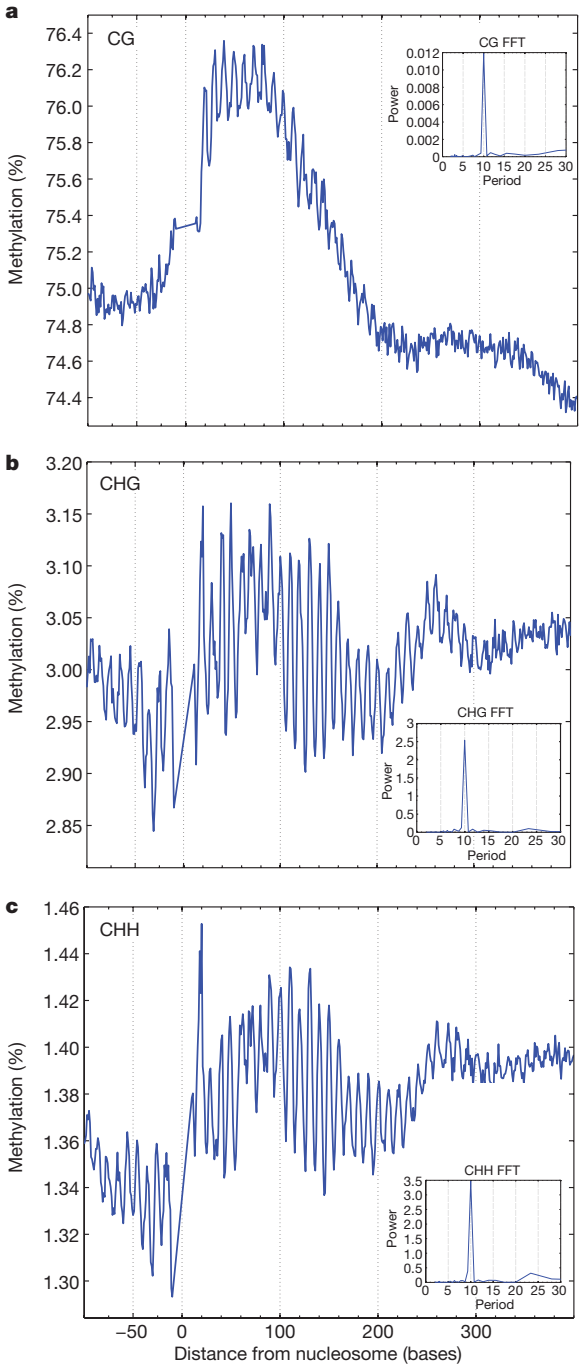


Figure 3 | DNA methylation profiles of nucleosome-bound DNA in humans. The weighted average percent DNA methylation was calculated and plotted at each distance from nucleosome start sites (0). **a**–**c**, CG methylation (**a**), CHG methylation (**b**) and CHH methylation (**c**) each show a 10-base periodicity over nucleosome-bound DNA (1–147 bases). The FFTs calculated over the region of the nucleosome (insets) demonstrate this periodicity in CG (**a**), CHG (**b**), and CHH (**c**) contexts.

analysed DNA methylation patterning on nucleosomes in different regions of the *Arabidopsis* and human genomes, including genes, promoters, pericentromeric regions and euchromatic arm regions (Supplementary Figs 9–17) and found that the 10-nucleotide periodicity was found in all cases, indicating that the relationship between nucleosome positioning and DNA methylation is general.

We observed that nucleosomes were much more abundant in exons than in introns (see Fig. 1a for an example), consistent with findings from several recent nucleosome positioning studies in other organisms[18–23]. The number of nucleosomes per base pair in introns was only 63% of the level found in exons. Furthermore, we found a strong peak of nucleosome start sites at intron–exon junctions and at exon–intron junctions (Fig. 4, Supplementary Fig. 18). The enrichment of nucleosomes in exons was not solely owing to higher G+C content (Supplementary Fig. 19a–c), or because of consensus splice site sequences (Supplemental Fig. 19d–f), and was confirmed using independent chromatin immunoprecipitation methods (Supplementary Fig. 20). Longer exons contained a higher number of

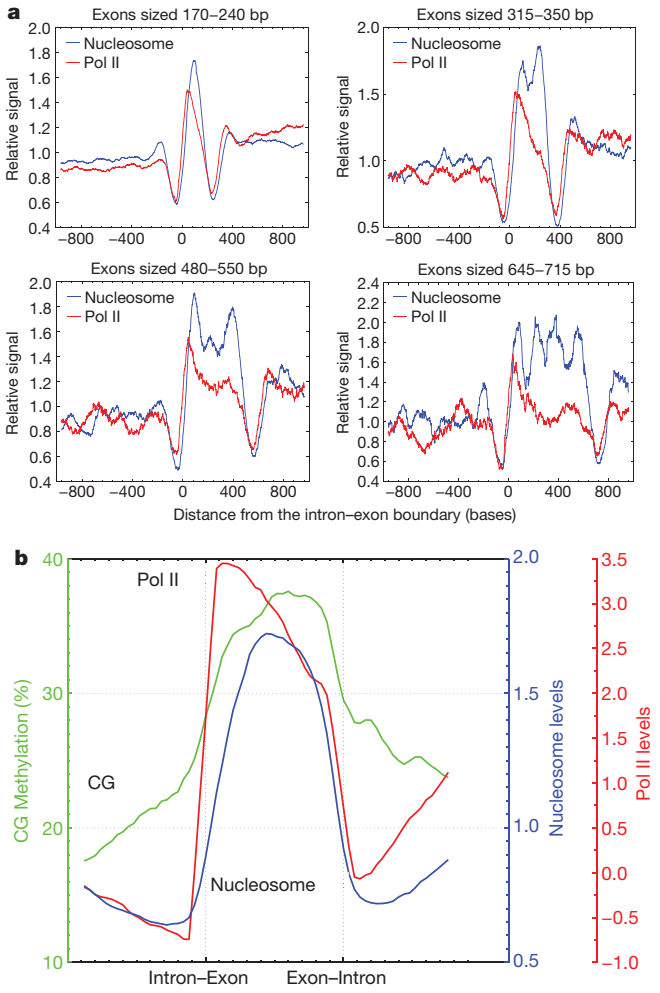


Figure 4 | Nucleosome and Pol II levels in exons. We performed chromatin immunoprecipitation with an antibody against RNA Pol II, and hybridized the resulting DNA to a whole genome Affymetrix microarray. We normalized these data to randomly sheared genomic DNA to control for probe efficiencies. **a**, Nucleosomes are phased in exons. Exon size limits were selected such that the exon could house no more than 1, 2, 3 or 4 nucleosomes. Nucleosome midpoints are plotted over these intron–exon boundaries. **b**, Each exon was divided into 25 equal sized bins and the nucleosome midpoints, Pol II and CG methylation levels are plotted over the exon and flanking introns.

nucleosomes. Examination of exons in the size ranges of 170–240, 315–350, 480–550 and 645–715 base pairs revealed peaks of one, two, three or four well-positioned nucleosomes respectively (Fig. 4a), indicating that nucleosomes are phased within exons.

Strong nucleosome enrichment on exons was detected both for genes that are highly expressed and for those not expressed (Supplementary Fig. 21 and Supplementary Table 1), suggesting that DNA sequences position nucleosomes in the absence of active transcription and splicing. Consistent with this hypothesis, using a nucleosome positioning prediction algorithm[24] we found similar patterns of nucleosome positioning in introns and exons between the experimental and theoretical data sets (Supplementary Fig. 22). Similarly, using theoretically predicted nucleosome positions, we observed similar patterns of DNA methylation on nucleosome-bound DNA (Supplementary Fig. 23), as well as enrichment of predicted nucleosomes in pericentromeric regions (Supplementary Fig. 24).

Because nucleosomes present a barrier to RNA Pol II transcription, we tested for Pol II occupancy in exons using a chromatin immunoprecipitation microarray approach. We observed significant enrichment of Pol II in exons relative to introns, consistent with the hypothesis that Pol II is paused on exonic DNA (Fig. 4). One possibility is that Pol II stalling on exons could enhance accurate splicing of upstream introns, thus reducing exon skipping and aiding in the fidelity of exon definition[20]. This is consistent with the finding of Pol II enrichment on human exons and indicates that Pol II enrichment on exons might be a common eukaryotic feature[20]. Furthermore, particular histone modifications have been recently shown to recruit splicing regulators, providing additional possible mechanisms for the regulation of splicing by nucleosome positioning[25].

Because of the enhancement of DNA methylation over nucleosomal DNA, and the enrichment of nucleosomes on exons, a prediction is that DNA methylation should be enhanced on exons relative to introns. Indeed, we found depletion of DNA methylation in introns and enrichment in exons in a pattern that was similar to nucleosome occupancy (Fig. 4b). This is consistent with recent findings of enhanced exonic methylation in other plant species, as well as in the *Chlamydomonas*, honeybee, *Ciona* and human genomes[26,27]. This suggests the possibility that DNA methylation, which frequently exists in the transcribed regions of active genes[26,28], could have a conserved role in exon definition or splicing regulation. These findings also reinforce the view that nucleosomal positions have an important role in shaping the methylation landscape of eukaryotic genomes.

METHODS SUMMARY

Approximately 300 ng of MNase-digested mononucleosome DNA were used for Illumina library generation following manufacturer instructions. Libraries were sequenced using an Illumina Genome Analyzer II following manufacturer instructions. Resulting reads were mapped to the TAIR7 annotation of the *Arabidopsis* genome and reads that mapped to multiple sites were eliminated. A total of 24.2 million reads were mapped to the forward strand and 23.9 million reads to the reverse strand, generating a 68-fold coverage of nucleosome space.

Pol II ChIP-chip was performed as described previously[29]. Briefly, crosslinked chromatin was extracted, sonicated and used in ChIP with an antibody against Pol II CTD (ab817; Abcam). Pol II-bound DNA and input genomic DNA were extracted, amplified, labelled and hybridized to Affymetrix tiling microarrays as described previously[29]. Three biological replicates were performed and the $\log_2$ ratios of Pol II-bound DNA over input DNA were calculated using the Tiling Analysis Software (Affymetrix).

Received 5 March; accepted 7 May 2010.
Published online 30 May 2010.

1. Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389,** 251–260 (1997).
2. Kornberg, R. D. & Lorch, Y. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* **98,** 285–294 (1999).
3. Cokus, S. J. *et al.* Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452,** 215–219 (2008).
4. Sun, F. L., Cuaycong, M. H. & Elgin, S. C. Long-range nucleosome ordering is associated with gene silencing in *Drosophila melanogaster* pericentric heterochromatin. *Mol. Cell. Biol.* **21,** 2867–2879 (2001).
5. Albert, I. *et al.* Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. *Nature* **446,** 572–576 (2007).
6. Mavrich, T. N. *et al.* A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res.* **18,** 1073–1083 (2008).
7. Valouev, A. *et al.* A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res.* **18,** 1051–1063 (2008).
8. Mavrich, T. N. *et al.* Nucleosome organization in the *Drosophila* genome. *Nature* **453,** 358–362 (2008).
9. Widom, J. Role of DNA sequence in nucleosome stability and dynamics. *Q. Rev. Biophys.* **34,** 269–324 (2001).
10. Segal, E. *et al.* A genomic code for nucleosome positioning. *Nature* **442,** 772–778 (2006).
11. Henderson, I. R. & Jacobsen, S. E. Epigenetic inheritance in plants. *Nature* **447,** 418–424 (2007).
12. Kobor, M. S. & Lorincz, M. C. H2A.Z and DNA methylation: irreconcilable differences. *Trends Biochem. Sci.* **34,** 158–161 (2009).
13. Jia, D., Jurkowska, R. Z., Zhang, X., Jeltsch, A. & Cheng, X. Structure of Dnmt3a bound to Dnmt3L suggests a model for *de novo* DNA methylation. *Nature* **449,** 248–251 (2007).
14. Lindroth, A. M. *et al.* Dual histone H3 methylation marks at lysines 9 and 27 required for interaction with *CHROMOMETHYLASE3*. *EMBO J.* **23,** 4146–4155 (2004).
15. Schones, D. E. *et al.* Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132,** 887–898 (2008).

16. Jeong, S. *et al.* Selective anchoring of DNA methyltransferases 3A and 3B to nucleosomes containing methylated DNA. *Mol. Cell. Biol.* **29,** 5366–5376 (2009).
17. Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Rev. Genet.* **11,** 204–220 (2010).
18. Andersson, R., Enroth, S., Rada-Iglesias, A., Wadelius, C. & Komorowski, J. Nucleosomes are well positioned in exons and carry characteristic histone modifications. *Genome Res.* **19,** 1732–1741 (2009).
19. Tilgner, H. *et al.* Nucleosome positioning as a determinant of exon recognition. *Nat. Struct. Mol. Biol.* **16,** 996–1001 (2009).
20. Schwartz, S., Meshorer, E. & Ast, G. Chromatin organization marks exon-intron structure. *Nat. Struct. Mol. Biol.* **16,** 990–995 (2009).
21. Kolasinska-Zwierz, P. *et al.* Differential chromatin marking of introns and expressed exons by H3K36me3. *Nature Genet.* **41,** 376–381 (2009).
22. Spies, N., Nielsen, C. B., Padgett, R. A. & Burge, C. B. Biased chromatin signatures around polyadenylation sites and exons. *Mol. Cell* **36,** 245–254 (2009).
23. Ponts, N. *et al.* Nucleosome landscape and control of transcription in the human malaria parasite. *Genome Res.* **20,** 228–238 (2010).
24. Kaplan, N. *et al.* The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458,** 362–366 (2009).
25. Luco, R. F. *et al.* Regulation of alternative splicing by histone modifications. *Science* **327,** 996–1000 (2010).
26. Feng, S. *et al.* Conservation and divergence of methylation patterning in plants and animals. *Proc. Natl Acad. Sci. USA* **107,** 8689–8694 (2010).
27. Laurent, L. *et al.* Dynamic changes in the human methylome during differentiation. *Genome Res.* **20,** 320–331 (2010).
28. Zemach, A., McDaniel, I. E., Silva, P. & Zilberman, D. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* advance online publication doi:10.1126/science.1186366 (15 April 2010).
29. Zhang, X. *et al.* Whole-genome analysis of histone H3 lysine 27 trimethylation in *Arabidopsis*. *PLoS Biol.* **5,** e129 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank members of our laboratories for input and guidance, and the Department of Energy, Institute for Genomics and Proteomics for support. R.K.C. is supported by NIH training grant GM07104. S.F. is a Howard Hughes Medical Institute Fellow of the Life Sciences Research Foundation. P.-Y.C. is funded by Eli & Edythe Broad Center of Regenerative Medicine & Stem Cell Research at UCLA. M.B. was supported by an Alexander-von-Humboldt Fellowship and U.K. by a Heisenberg Fellowship. X.Z. was supported by a Faculty Research Grant (JR-040) from the University of Georgia. Research was supported by an Innovation Award from the Eli & Edythe Broad Center of Regenerative Medicine & Stem Cell Research at UCLA, NIH grants GM60398 and GM42143, and NSF Plant Genome Research Program #0701745. S.E.J. is an Investigator of the Howard Hughes Medical Institute.

Author Contributions S.E.J. and M.P. designed the research. S.F., Y.V.B., H.S., Y.Y., J.A.H., M.B., P.H., A.T.C., U.K., S.S.M. and X.Z. performed experiments. R.K.C., P.-Y.C, F.K., J.K., S.J.C and D.C. analysed data. R.K.C., S. F., S.E.J. and M.P. wrote the paper.

Author Information All sequencing files have been deposited in GEO under accession code GSE21673. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to S.E.J. (jacobsen@ucla.edu) or M.P. (matteop@mcdb.ucla.edu).