Title

Studies on the Interaction of DGCR8 with the Cofactor Heme.

Author

Barr, Robert Ian

Publication Date

2012

UNIVERSITY OF CALIFORNIA
Los Angeles

Studies on the Interaction of DGCR8 with the Cofactor Heme

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Biological Chemistry

by

Robert Ian Barr

2012

ABSTRACT OF THE DISSERTATION

Studies on the Interaction of DGCR8 with the Cofactor Heme

by

Robert Ian Barr

Doctor of Philosophy in Biological Chemistry

University of California, Los Angeles, 2012

Professor Feng Guo, Chair

The RNA-binding protein DiGeorge Critical Region 8 (DGCR8) and its partner, the nuclease Drosha, are necessary for processing of primary microRNA (pri-miRNA) transcripts in animals. MicroRNAs, in turn, are essential pieces of the cell's regulatory machinery, where they serve to post-transcriptionally regulate gene expression through the RNA-Induced Silencing Complex. Work described in this dissertation concerns the interaction of DGCR8 with the cofactor heme, and shows how the activity of DGCR8 — and by extension, the activity of Drosha — is controlled by the presence of heme within the heme-binding domain. Biochemical and spectrophotometric assays are used to show that the heme in DGCR8 is in the Fe(III) form and is ligated to two cysteines, one from each subunit of the DGCR8 homodimer. The activity of DGCR8 is shown to depend on the oxidation state of the iron; Fe(II) is unable to activate processing or allow trimerization of DGCR8 on pri-miRNA. I also show that, from a screen of ten metalloporphyrins, only Co(III) protoporphyrin IX is able to both bind to DGCR8 and activate pri-miRNA processing. The implications of this work are discussed.

The dissertation of Robert Ian Barr is approved.

Guillaume Chanfreau

Reid Johnson

David S. Eisenberg

Feng Guo, Committee Chair

University of California, Los Angeles
2012

Dedicated to my family, and to

Christine, who helped me see it through.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

My co-authors, Rachel Senturia, Yanqiu Chen, Brooke Scheidemantle, Feng Guo, Aaron Smith, and Judith Burstyn made this work possible. Michael Faller started this project and did a lot of the difficult, early work on heme-bound DGCR8. We thank Juli Feigon for sharing the spectrophotometer in her lab and Jose Paul Jacob for technical assistance. I'd like to thank Martin Phillips from the Biochemistry Instrumentation Facility at UCLA for his advice and training. I'd like to acknowledge support from the Audree Fowler Fellowship in Protein Science, and the UCLA Dissertation Year Fellowship. This project was supported by a grant from the National Institutes of Health (GM080563) to Feng Guo.

VITA

1999–2004	B.S. (Biochemistry), UCSB, Santa Barbara, California
2004–2006	M.S. (Chemistry), Portland State University, Portland, Oregon.

PUBLICATIONS

Barr I., Guo F. (2012) DGCR8 is activated for primary-microRNA processing by Co(III) Protoporphyrin IX. (In preparation; Chapter 4 in this volume.)

Barr I., Guo F. (2012) *In vitro* Primary-microRNA Processing using Recombinant DGCR8 and Drosha. *Methods in Molecular Biology*, in press.

Senturia R., Laganowsky A., **Barr I.**, Scheidemantle B.D., Guo F. (2012) Dimerization and Heme Binding Are Conserved in Amphibian and Starfish Homologues of the microRNA Processing Protein DGCR8. *PLoS ONE*, 7, e39688.

Meer E. J., Ohtan Wang D., Kim S. M., **Barr I.**, Guo F., and Martin K. C. (2012) Identification of a *cis*-acting element that localizes mRNA to synapses. *Proc. Natl. Acad. Sci. USA*, 109, 4639-4644.

Barr I. , Smith A. T. , Chen Y. , Senturia R. , Burstyn J. N. , Guo F (2012) Ferric, Not Ferrous, Heme Activates DGCR8 for Primary microRNA Processing. *Proc. Natl. Acad. Sci. USA*, 109, 1919-1924.

Barr I. , Smith A. T. , Senturia R. , Chen Y. , Scheidemantle B. D. , Burstyn J. N. , Guo F (2011) DiGeorge Critical Region 8 (DGCR8) is a double-cysteine ligated heme protein. *Journal of Biological Chemistry* 286, 16716-16725.

Wang Q. , **Barr I.** , Guo F. , Lee C. (2008) Evidence of a novel RNA secondary structure in the coding region of HIV-1 pol gene. *RNA* 14, 2478-2488.

CHAPTER 1

Introduction

DGCR8 is Required for miRNA Processing in Animals.

MiRNAs are involved in the regulation of gene expression in animals and plants [1]. In humans, over 400 miRNAs have been identified to date [2]. They bind to target mRNAs using a short seed region typically spanning positions 2-7 from the $5'$ end of a mature miRNA that is complementary to the mRNA. MiRNAs are transcribed in the nucleus by RNA polymerases II and in some cases possibly RNA polymerase III [3, 4]. They are often in complexes with many miRNAs in one polycistronic transcript [5, 6, 7]. They can be located in their own loci, or be placed within the introns of other genes [8]. Once transcribed, pri-miRNAs are processed by a protein complex called the Microprocessor into $\sim$ 65 nt hairpin RNAs called precursor microRNAs (pre-miRNAs) (Figure 1.1). Pre-miRNAs are then exported from the nucleus by Exportin-5 and cleaved by the nuclease Dicer. Dicer helps load the mature microRNAs into RISC, the **R**NA-**I**nduced **S**ilencing **C**omplex, where they serve to recognize mRNAs and regulate them. The Argonaute family of proteins [9] is the central component of RISC responsible for forming a complex with the mature miRNA and recognizing the mRNA substrate. In mice and humans, Argonaute 2 (Ago-2) is the only one of the four Agos that is able to catalyze cleavage of the target mRNA. Estimates of the number of protein-coding genes regulated by the miRNA machinery range from 20-80%.

The cleavage step where pri-miRNA is processed into pre-miRNA is carried out by the RNase III family member Drosha (Figure 1.2). Human Drosha is located on

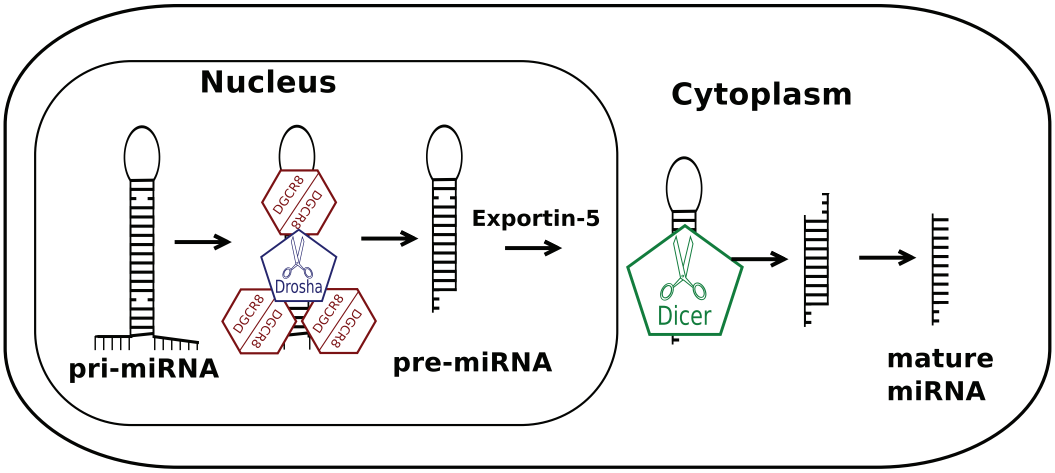

Figure 1.1: **Diagram of the microRNA Biogenesis Pathway in animals.**

human chromosome 5, and chromosome 15 in mice. In mice, it is expressed most highly in the brain, heart, kidneys, and testes, and less in skeletal muscle [10]. Drosha was originally thought to be involved in the maturation of ribosomal RNA, like its yeast homologues [11]; however, subsequent work showed that it was necessary for the cleavage of miRNAs [12].

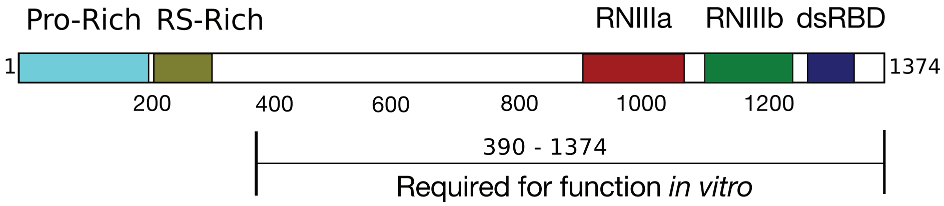

Figure 1.2: **Domain structure of human Drosha.**

DGCR8, called Pasha in *Drosophila* species and Pash-1 in *C. elegens*, was found to be necessary for pri-miRNA processing as well [13, 14, 15, 16]. DGCR8 stands for DiGeorge Syndrome critical region 8, which is derived from the fact that DGCR8 is one of around 30 genes deleted in people with DiGeorge Sydrome, a developmental disorder caused by deletion of a section of chromosome 22 [17]. Deficiency of DGCR8 has been shown to present some of the symptoms of DiGeorge syndrome [18, 19, 20]. Analysis of its expression profile showed it was expressed similarly to Drosha, with high expression in the brain and heart [17]. Although DGCR8 is known to associate with several other proteins [16, 21], DGCR8 and Drosha are sufficient to reconstitute the cleavage of pri-miRNA *in vitro* [15, 16].

There are two competing models of pri-miRNA recognition by the Microprocessor. In the first [22, 23], DGCR8 recognizes the loop at the end of the pri-miRNA stem, and the correct cleavage distance is measured from there. In the other [24], DGCR8 recognizes the stem loop junction at the other side and measures from there to determine the correct length. However, recent work has shown that DGCR8 recognizes pri-miRNA by binding as a cooperative trimer of homodimers (Figure 1.3); the ability to trimerize was shown to be correlated with processing by Drosha [25]. Neither of the above models takes into account the observed fact that DGCR8 forms a multimeric complex on pri-miRNA; the assumption is that there is one DGCR8 for one pri-miRNA. The size of the multimeric DGCR8 complex is sufficient for it to recognize both sides of the stem loop, allowing a more comprehensive model than previously imagined.

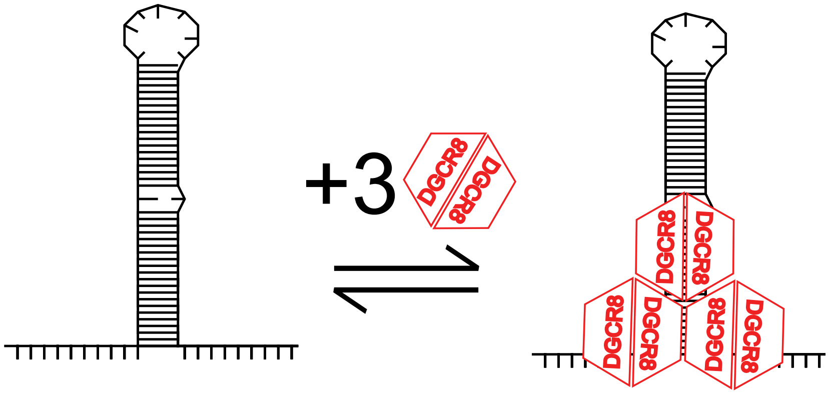

Figure 1.3: **Model of how DGCR8 Homodimers recognize pri-miRNA.**

Sequence analysis of DGCR8 shows that it contains, as one might expect, domains involved in interaction with RNA. Specifically, there are two canonical double-stranded RNA-binding domains (dsRBDs) located in the C-terminal region of the protein. Apart from these, the only region detected by early researchers was a WW motif in the region around a.a. 300 [16].

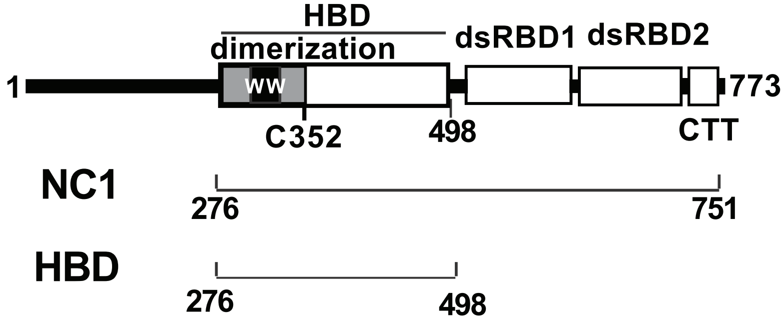

Figure 1.4: **Domain structure of human DGCR8.**

A crystal structure of the dsRBD region (a.a. 493-720) by Sohn et al [26] showed that the two dsRBDs were structurally similar to other known dsRBDs, with an additional helix on each. The two dsRBDs show pseudo two-fold symmetry in the crystal structure. Another crystal structure by Rachel Senturia [27] showed that the WW region of DGCR8 is dissimilar to all other known WW domains in that it forms the center of a dimerization structure, responsible in part or in whole for forming DGCR8 into a homodimer.

DGCR8 is a Hemoprotein.

Michael Faller in the Guo lab discovered that DGCR8 associates with heme, and this work forms the starting point of most of the research presented in this thesis [28]. He found, specifically, that a construct of human DGCR8 containing amino acids 276–751 was associated with heme B when expressed recombinantly in *E. coli*; however, bacteria only produce heme B and heme O, and would not be expected to contain the same heme transport machinery that animals do. Later, expression of NC1 in Sf9 insect cells gave the same absorbance spectrum (see **Chapter 2**).

The heme in DGCR8 has characteristic absorbance bands at 367, 450, and 556 nm (Figure 1.5; Table A.2). At first, this result was puzzling due to the fact that no known heme protein naturally absorbed at these particular wavelengths; however, see **Chapter**

2 for details about how this issue was resolved. In short, it was suspected that DGCR8 was bound to ferrous (Fe(II)) heme, based on the similarity of the peak to that seen in the Fe(II) form of Cytochrome P450, which also absorbs at 450 nm. Faller's work also showed that mutation of the conserved cysteine 352 led to loss of dimerization and loss of heme binding ability. Mass spectrometry and analytical chromatography showed that DGCR8 binds heme with a 2:1 protein subunit:heme ratio.

In the initial report, the role of heme was not determined. It was speculated that heme might act as a signaling molecule, activating or inactivating DGCR8 according to the current cellular heme concentration. Another option was that it could be a detector for a small molecule, possibly controlling pri-miRNA processing through binding to a gas such as carbon monoxide. Using these facts as a starting point, we set out to piece together the full story. This required familiarizing ourselves with the ancient and well-traveled field of heme chemistry.

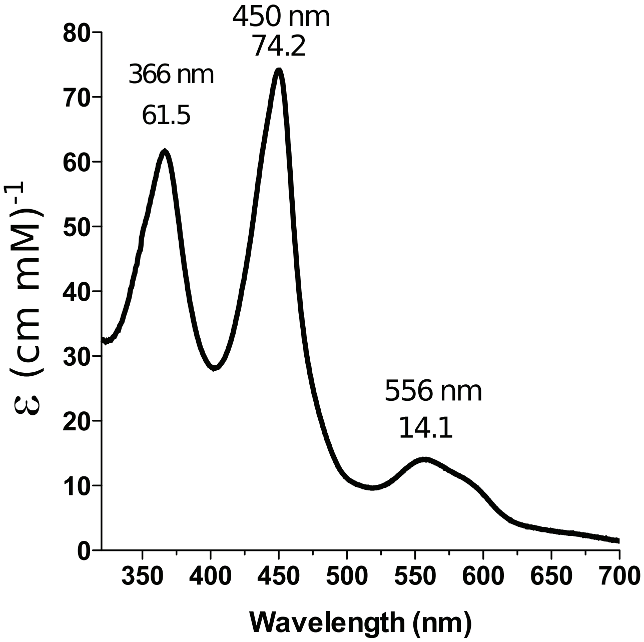

Figure 1.5: **Electronic Absorbance Spectrum of DGCR8 NC1, with extinction coefficients.**

Heme interactions with proteins.

The heme molecule and its interactions with proteins have been studied for over 100 years [29, 30]. As early as 1863, G.G. Stokes was monitoring changes in the hemoglobin absorbance spectrum upon reduction of the heme to the Fe(II) form [31].

> I had no sooner looked at the spectrum, than the extreme sharpness and beauty of the absorption-bands of blood excited a lively interest in my mind, and I proceeded to try the effect of various reagents. The observation is perfectly simple, since nothing more is required than to place the solution to be tried, which may be contained in a test-tube, behind a slit, and view it through a prism applied to the eye.

Heme-binding proteins were also among the first proteins purified and characterized. A major factor in this was their strong absorbance in the visible region, noted by Stokes in the above quote, which allowed detection by eye using primitive spectroscopic techniques. This strong absorbance is largely due to the highly conjugated tetrapyrrole ring they contain, giving extinction coefficients at their strongest band (the Soret band) on the order of 100,000 $cm^{-1}M^{-1}$ (see **Appendix Table A.1** for examples). From the very beginning there was a strong link between heme chemistry and spectroscopy. The physiologist and chemist Karl von Vierordt used absorbance to determine the concentration of hemoglobin in 1872, at the same time as he made one of the first successful applications of the Beer-Lambert law [32, 33]. Biochemists found heme derivatives in urine and feces using spectroscopy [34, 35]. Keilin [36] was able to identify three different species of cytochrome, *a*, *b*, and *c*, using characteristic absorbance peaks in the 400–800 nm region. Kurt Stern [37, 38] detected reaction intermediates in catalase using the heme α and β absorbance bands. As you will see from my work, this connection remains to this day an intimate one.

Heme proteins are also commonly studied by electron paramagnetic resonance (EPR) [39, 40], resonance Raman (rR), Fluorescence [41], and Magnetic Circular Dicroism (MCD) Spectroscopy [42]. Each of these techniques is normally viewed as complimentary to the others; so it is quite common to use as many as possible when characterizing a new system. The spectroscopists are normally looking for information about the chemical environment around the heme, the metal's spin and charge states, the redox potential, and the identity of the ligands to the heme. As we show in **Chapter 2**, several spectroscopic techniques – MCD, EPR, and electronic absorbance – were necessary to determine the native state of the heme in DGCR8.

In addition to the above methods, there exist several common ways of determining the dynamic and function aspects of the heme in a hemoprotein. Each hemoprotein has a dissociation constant K_D, giving an idea how strongly the heme associates with the protein. If the apoprotein can be easily made, one may be able to determine the association and dissociation rates of heme. Association rates for hemoproteins can be very fast, so stopped-flow or similar methods are needed. In **Chapter 5** I report the results of work along this line.

Heme can have several oxidation states, depending upon its environment. Most commonly, it is in the Fe(II) and Fe(III) form, though other states are possible (see [30]).The fact that heme is redox active opens up another area of research; it is, in fact, the subject of the paper in **Chapter 4**. The study of redox transitions in metalloproteins using ultra-fast kinetics [43, 44] gives us an idea of how redox potential is utilized by cells to control the flow of electrons through proteins. Spectroscopists are able to detect the formation of kinetic intermediates in enzymatic hemes [45]. The field has, as a whole, benefited from the interest of physical chemists and inorganic chemists who have brought many of these techniques along with them into the heme field. In **Chapter 3** we combine biochemical and spectroscopic techniques to determine whether, for DGCR8, the heme oxidation state makes a difference for its activity. Heme is synthesized in the Fe(II) form by Ferrochelatase, the final enzyme in the heme biosynthesis pathway. However, both the Fe(II) and Fe(III) states are accessible at physiological redox potentials. Our research suggests that, for DGCR8, the correct redox state of the heme is crucial for pri-miRNA binding activity and subsequent cleavage by Drosha.

In the **Appendix**, I describe several of the analytical techniques used in the lab: how they were developed, how we use them, and how the necessary equations used in analysis were derived. I also give the formulas for reagents used in my work.

CHAPTER 2

DGCR8 is a double-cysteine ligated heme protein

(*Journal of Biological Chemistry* 286, 16716-16725.)

DiGeorge Critical Region 8 (DGCR8) Is a Double-cysteine-ligated Heme Protein

Received for publication, September 13, 2010, and in revised form, March 18, 2011 Published, JBC Papers in Press, March 21, 2011, DOI 10.1074/jbc.M110.180844

Ian Barr[‡], Aaron T. Smith[§], Rachel Senturia[‡], Yanqiu Chen[‡], Brooke D. Scheidemantle[‡], Judith N. Burstyn[§1], and Feng Guo[‡2]

From the [‡]Department of Biological Chemistry, David Geffen School of Medicine, Molecular Biology Institute, University of California, Los Angeles, California 90095, and the [§]Department of Chemistry, University of Wisconsin, Madison, Wisconsin 53706

All known heme-thiolate proteins ligate the heme iron using one cysteine side chain. We previously found that DiGeorge Critical Region 8 (DGCR8), an essential microRNA processing factor, associates with heme of unknown redox state when overexpressed in *Escherichia coli*. On the basis of the similarity of the 450-nm Soret absorption peak of the DGCR8-heme complex to that of cytochrome P450 containing ferrous heme with CO bound, we identified cysteine 352 as a probable axial ligand in DGCR8. Here we further characterize the DGCR8-heme interaction using biochemical and spectroscopic methods. The DGCR8-heme complex is highly stable, with a half-life exceeding 4 days. Mutation of the conserved proline 351 to an alanine increases the rate of heme dissociation and allows the DGCR8-heme complex to be reconstituted biochemically. Surprisingly, DGCR8 binds ferric heme without CO to generate a hyperporphyrin spectrum. The electronic absorption, magnetic circular dichroism, and electron paramagnetic resonance spectra of the DGCR8-heme complex suggest a ferric heme bearing two cysteine ligands. This model was further confirmed using selenomethionine-substituted DGCR8 and mercury titration. DGCR8 is the first example of a heme-binding protein with two endogenous cysteine side chains serving as axial ligands. We further show that native DGCR8 binds heme when expressed in eukaryotic cells. This study provides a chemical basis for understanding the function of the DGCR8-heme interaction in microRNA maturation.

Heme proteins use a variety of amino acids to serve as the axial ligands (the fifth and possibly sixth ligands) for the heme iron. In many cases, such as with hemoglobins (1), soluble guanylate cyclases (2) and heme oxygenases (3), an imidazole moiety of a histidine side chain acts as the fifth ligand. Another major class of heme-binding proteins, exemplified by the cytochrome P450 proteins (4), chloroperoxidases (5), and nitric oxide synthases (6), uses a cysteine side chain as the fifth ligand. This class is called heme-thiolate proteins (7). The functions of these axial ligands include enhancing the affinity and specificity of the protein-heme interaction, modulating the redox properties of the heme, and enabling the heme to serve as catalysts, transporters, or sensors (8, 9). Often, the sixth ligand of the iron in heme is a substrate, a cargo, or a protein moiety that modulates the biological activities of the protein.

We have previously found that the RNA-binding partner protein of the Drosha nuclease in humans, called DiGeorge Critical Region 8 (DGCR8)[3] (or Pasha for its insect and worm homologs), is a heme-binding protein (10). DGCR8 (11–14) and Drosha (15) are responsible for processing primary transcripts of microRNAs (pri-miRNAs) into precursor miRNAs in the first step of the miRNA maturation pathway in animals (16–18). A truncated but active form of DGCR8 named NC1, containing residues 276–751 of this 773-amino acid protein (Fig. 1*A*), exists as a heme-bound dimer and heme-free monomer when overexpressed in *Escherichia coli* (10). The heme-bound NC1 dimer is more active in reconstituted pri-miRNA processing assays than the heme-free monomer, indicating that heme is important for the function of DGCR8. Heme-binding-deficient NC1 mutants are defective in pri-miRNA processing *in vitro* (10, 19). DGCR8 binds heme through a heme-binding domain (HBD, amino acids 276–498) that can be expressed in a soluble, heme-bound form in the absence of other domains (Fig. 1*A*) (19). The HBD is not required for binding and processing pri-miRNAs *in vitro* and most likely serves a regulatory function (10). The N-terminal region of the HBD (amino acids 276–353) is a dimerization subdomain that includes a WW motif, which is most often involved in binding proteins with proline-containing sequences (19). The crystal structure of this domain, in the absence of heme, revealed a novel use of its WW motif for dimerization (19). Importantly, the structure showed that all the residues known to be essential for association with heme, including Cys-352 and Trp-329 (the second W in the WW motif) from both subunits, cluster on a common surface that presumably directly contacts heme.

---

[3] The abbreviations used are: DGCR8, DiGeorge Critical Region 8; miRNA, microRNA; pri-miRNA, primary transcripts of microRNA; HBD, heme-binding domain; δ-ALA, δ-aminolevulinic acid; SEC, size exclusion chromatography; hHBD^{C430S}, human DGCR8 HBD-His$_6$ C430S mutant protein; EPPS, 4-(2-hydroxyethyl)-1-piperazinepropanesulfonic acid; ϵ, extinction coefficient; MCD, magnetic circular dichroism; MeHgAc, methylmercury acetate; SeMet, selenomethionine.

DGCR8 appears to belong to a previously uncharacterized family of heme proteins. Heme-bound NC1 exhibits electronic absorption peaks at 450, 367, and 556 nm, with the 367-nm peak almost as intense as the strongest 450-nm peak. The Soret peak at ~450 nm is characteristic for heme-binding proteins with a thiolate ligand contributed by a cysteine side chain (20, 21). The Soret peak of the heme-thiolate protein cytochrome P450 red-shifts to ~450 nm when the iron in heme is reduced to the ferrous state and CO binds as the sixth ligand (22). Consistent with this precedent, mutation of a conserved Cys-352 to alanine, serine, or histidine in NC1 abolishes heme binding, implying that Cys-352 is an axial ligand of the iron in the heme (10). However, there is no obvious source of CO in the purification and storage of recombinant NC1. Furthermore, in cytochrome P450, ferrous heme ligated by a cysteine thiolate *trans* to CO exhibits a spectrum with the 366-nm band substantially less intense than the ~450-nm peak. Our understanding of the DGCR8-heme interaction has been limited by a lack of *in vitro* heme-binding assays, as incubation of the heme-free monomer with heme does not produce the heme-bound dimer. In addition, there does not appear to be any interchange between the dimer and monomer species *in vitro*, regardless whether heme is present.

In this study, we used a diverse set of methods to examine the redox and ligation states of the heme in DGCR8. A mutant NC1, P351A, was shown to reversibly bind heme, forming a reconstituted complex that displays an electronic absorption spectrum similar to that of the wild-type protein. Heme dissociation and association assays revealed heretofore unknown properties of DGCR8. Most notably, DGCR8 stably binds ferric heme and exhibits the 450-nm Soret peak in the absence of CO. Spectroscopic characterization of the DGCR8 HBD reveals unique features associated with the uncommon heme coordination environment, including a narrow *g* anisotropy in the EPR spectrum. These results, along with SeMet substitution and mercury titration experiments, demonstrate that DGCR8 is a novel heme-binding protein with two cysteine side chains as axial ligands. Additionally, we show that DGCR8 binds heme when expressed in eukaryotic cells. The unique interaction between DGCR8 and heme provides the biochemical basis for understanding a potential mechanism for regulating miRNA maturation.

EXPERIMENTAL PROCEDURES

Reagents—Horse skeletal muscle apomyoglobin was from Sigma-Aldrich (St. Louis, MO). Hemin and δ-ALA were from MP Biomedicals (Solon, OH). L-(+)-selenomethionine was from Acros Organics (Morris Plains, NJ). Methylmercury(II) acetate (MeHgAc) was purchased as a part of a heavy atom screening kit from Hampton Research (Aliso Viejo, CA). Caution: MeHgAc is extremely toxic and should be handled with great care.

Plasmids—Site-directed mutagenesis of NC1 and human DGCR8 HBD-His$_6$ in the pET-24a+ vector was carried out using the standard four-primer PCR method (19). The insect cell expression plasmid for His$_6$-NC1 was generated by cloning the coding sequence for residues 276–751 of human DGCR8 into the pFastBac-HTb vector between the BamHI and EcoRI sites. The cDNA clone of the DGCR8 homolog from *Xenopus laevis* (frog) was obtained from Open Biosystems. The HBD coding region of the frog DGCR8 (amino acids 278–498) was amplified using PCR and was inserted into the pET-24a+ vector between the NdeI and NotI sites. The coding sequences of all plasmids were verified through sequencing.

Protein Expression and Purification—NC1 constructs were expressed and purified as described previously (10). The heme-bound proteins were expressed with 1 mM δ-ALA added at induction. The purification procedure started with two rounds of ion exchange chromatography using a SP-Sepharose HP column, and the monomer and dimer species were separated using size exclusion chromatography (SEC) and a Superdex 200 10/300 GL column (GE Healthcare).

The frog DGCR8 HBD-His$_6$ was expressed in *E. coli* strain BL21(DE3) CodonPlus (Strategene) and purified using Ni-affinity chromatography followed by SEC, as reported previously for human HBD-His$_6$ (19). The purified protein was exchanged into desired buffers for spectroscopic studies, as indicated in the figure legends, using centrifugal concentrators with a molecular mass cut-off of 30 kDa. SeMet-labeled frog DGCR8 HBD-His$_6$ was expressed using the same vector, host strain, and procedure as described by Van Duyne *et al.* (23). Briefly, *E. coli* were cultured in M9 minimal medium containing no methionine. Six other amino acids known to inhibit methionine biosynthesis were added to the culture 15 min prior to induction. In addition, 40 μM $FeCl_3$ and 1 mM δ-ALA were added to the medium at induction to increase the synthesis of heme and the yield of the HBD-heme complex. SeMet-labeled HBD-His$_6$ was expressed at 25 °C overnight (~16 h) and purified using the same procedure as that for the unlabeled protein.

The human DGCR8 HBD-His$_6$ C430S mutant protein (hHBD^{C430S}, used in a mercury titration experiment) was expressed and purified using the same procedure as described for the wild-type protein (19), except that the SEC was performed in the absence of reducing reagents. Prior to the mercury titration experiments, the purified HBD mutant was concentrated and exchanged into a buffer containing 50 mM EPPS (pH 8.0) and 400 mM NaCl, using a centrifugal concentrator with a molecular mass cut-off of 30 kDa. The extinction coefficient of the heme-bound hHBD^{C430S} protein ($\epsilon_{280\ nm}$ = 31 $mM^{-1}\ cm^{-1}$) was determined experimentally by measuring its absorption at 280 nm and its protein concentration using the Micro BCA protein assay kit (Pierce).

The His$_6$-NC1 protein was expressed in Sf9 cells following the instructions in the Bac-to-Bac system (Invitrogen). His$_6$-NC1 was purified using Ni-affinity chromatography followed by cation exchange chromatography using a SP column (GE Healthcare). The purified protein fractions were pooled, concentrated, and exchanged in a storage buffer containing 20 mM Tris (pH 8.0), 400 mM NaCl, and 1 mM DTT using an Ultra-4 concentrator with a 30-kDa molecular mass cut-off. Sf9 cell-expressed His$_6$-NC1 was analyzed using a Superdex 200 10/300 GL column, and the absorption at 260, 280, and 450 nm was monitored (Fig. 9*B*). The size exclusion chromatogram was baseline-adjusted using the Unicorn software (version 5.0, GE Healthcare). The ratio of the absorption by heme and by protein (the *RZ* value) was used to calculate the percentage of His$_6$-NC1 occupied by heme using the extinction coefficients (ϵ) 58 mM^{-1} cm^{-1} for heme (450 nm) and 53 mM^{-1} cm^{-1} for each His$_6$-NC1 subunit (280 nm). The $\epsilon_{450\text{ nm}}$ of insect cell-expressed His$_6$-NC1 was assumed to be the same as that of the untagged NC1 expressed in *E. coli* (10). The $\epsilon_{280\text{ nm}}$ was calculated from the His$_6$-NC1 sequence using the program ProtParam on the ExPASy proteomics server provided by the Swiss Institute of Bioinformatics (24).

Heme Dissociation and Association Assays—Hemin chloride was dissolved in 1.4 M NaOH at 100 mM concentration and incubated at room temperature for >30 min. This concentrated solution was then diluted in water to give a 100 μM stock solution. All reactions were carried out in 20 mM Tris buffer (pH 8.0), 400 mM NaCl, and 1 mM DTT, unless otherwise specified.

Electronic Absorption Spectroscopy—The spectra were recorded at 25 °C using either a DU800 spectrophotometer (Beckman-Coulter, bandwidth $\leq$ 1.8 nm) in sealed cuvettes (Figs. 1–4, 8, 9), a DU500 spectrophotometer (Beckman-Coulter, bandwidth < 4.5 nm) (Fig. 3*D*), or a double-beam Varian Cary 4 Bio spectrophotometer (Agilent Technologies) with its spectral bandwidth set to 0.5 nm (Figs. 5 and 7*B*).

MCD Spectroscopy—MCD spectra were recorded on a Jasco J-715 circular dichroism spectropolarimeter with the sample compartment modified to accommodate an SM-4000–8T magnetocryostat (Oxford Instruments). The buffer used for the MCD samples was 50 mM EPPS and 400 mM NaCl (pH 8.0) with ~55% (v/v) glycerol present in the final sample. Glycerol was introduced to the protein and stirred with a syringe until the solution was homogenous. Glycerol had no effect on the electronic absorption spectra at room or liquid-helium temperatures. Samples were transferred via gas-tight syringes into cells, flash-frozen, and stored in liquid N$_2$. MCD spectra were taken over a temperature range from 4 to 50 K. The MCD signal at each temperature was recorded at ±7 T. Negative polarity data were subtracted from positive polarity data to remove circular dichroism contributions, and the resulting spectrum was divided by 2. Magnetic saturation curves were generated from data taken at 448 nm, and data were normalized to the most intense MCD data point (7 T, 2.5 K).

EPR Spectroscopy—X-band electron paramagnetic resonance spectra were collected on a Bruker ELEXSYS E500 equipped with an Oxford ESR 900 continuous flow liquid helium cryostat and an Oxford ITC4 temperature controller maintained at 10 K. The microwave frequency was monitored using an EIP model 625A CW microwave frequency counter. Protein samples were prepared as indicated in the figure legends. Samples were transferred to a quartz EPR tube via small-bore tubing connected to a gas-tight syringe and were frozen in liquid N$_2$. Samples of 150 μl had a final concentration of ~150 μM heme. For all samples, scans of 0–10,000 G revealed no signals other than those reported.

MALDI-TOF Mass Spectrometry—Native and SeMet-labeled frog DGCR8 HBD-His$_6$ were dissolved in a 1:9 (v/v) ratio with a mixture containing 70% (v/v) acetonitrile, 0.1% (v/v) trifluoroacetic acid, and saturating concentration of sinapinic acid (Fluka). The samples were analyzed using a Voyager-DE STR MALDI-TOF mass spectrometer (Applied Biosystems) operated in a linear mode with positive polarity. The mass accuracy in this mode is $\leq \pm$ 0.05% for myoglobin (16,952 Da). The *m/z* range from 10,000 to 70,000 was monitored.

Mercury Titration of Cysteines in Human HBD—A stock solution of MeHgAc was prepared at a 0.5 mM concentration in 50 mM EPPS (pH 8.0) and 400 mM NaCl. (Caution: MeHgAc is volatile and extremely toxic. The stock solution was prepared in a fume hood.) The MeHgAc stock was added to 200 μl of 10-μM hHBD^{C430S} dimer solution in 2-μl increments. The solution was incubated at 25 °C for 10 min before the electronic absorption spectra were recorded. After the mercury titration, 30 μl of apomyoglobin stock solution (80 μM) was added, and the electronic absorption spectrum was recorded.

Reconstituted pri-miRNA Processing Assays—The procedure was described previously (10). Briefly, the 150-nucleotide pri-miR-30a was uniformly labeled with ^{32}P and was incubated with recombinant Drosha and His$_6$-NC1 at 37 °C for 45 min. The reactions were analyzed using denaturing 15% polyacrylamide gels and autoradiography.

RESULTS

The DGCR8 P351A Mutant Can Be Expressed as Heme-bound and Heme-free Dimers—Pro-351 is located close to the proposed heme-binding surface of DGCR8 in three-dimensional space (19). We mutated Pro-351 to alanine in the context of NC1 (P351A) and examined the ability of this mutant to bind heme by expressing it in *E. coli* with and without the addition of δ-ALA, a heme biosynthesis intermediate. In both cases, P351A was present in dimeric and monomeric forms, as indicated by SEC (Fig. 1, *B* and *C*). When overexpressed without δ-ALA, P351A was nearly completely free of heme (Fig. 1*B*). When δ-ALA was added to the bacterial culture at the time of induction, the heme occupancy of the P351A dimer was increased to nearly the level of the wild type under the same conditions; however, the P351A monomer remained free of heme (Fig. 1*C*). The heme-bound P351A dimer, purified using SEC, exhibited an electronic absorption spectrum with peaks at 365, 447, and 556 nm, similar to those of the wild-type NC1 protein (Fig. 1*C*). The P351A mutation caused a 3-nm blue shift of the 450-nm Soret peak, indicating some alteration in the local structure of the NC1 protein. In the presence of δ-ALA, the wild-type NC1 is expressed nearly exclusively as a heme-bound dimer (10), suggesting that the P351A mutation may reduce the affinity of DGCR8 for heme. Overall, these results suggest that the complex of P351A with heme retains a structure similar to that of the wild type.

Heme Dissociates from Wild-type DGCR8 Extremely Slowly, and the P351A Mutation Increases the Dissociation Rate—To determine the stability of the DGCR8-heme complex, we performed a heme dissociation assay by incubating heme-bound NC1 with 5-fold molar excess of horse apomyoglobin. Apomyoglobin has an extremely high affinity for heme, with K_D of ~3×10^{-15} M (25). Thus, heme molecules dissociated from NC1 were quickly bound to apomyoglobin, and this transfer was accompanied by a shift of the Soret peak from ~450 to 414 nm (Fig. 2*A*). At room temperature, heme dissociated from the P351A dimer within 1–2 h (Fig. 2*A*). When the absorbance at 447 nm was plotted over time, the data fit nicely to a one-phase exponential decay function, giving an estimated dissociation rate constant (k_{off}) of $5.0 \pm 0.1 \times 10^{-4}$ s^{-1}. The increase in absorbance at 414 nm over time gave a similar value ($5.8 \pm 0.4 \times 10^{-4}$ s^{-1}). In contrast, incubation of wild-type heme-bound NC1 with apomyoglobin did not result in any shift of the Soret peak even after 4 days (Fig. 2*C*). This observation suggests either extremely slow dissociation of heme from the wild-type NC1 (with $t_{1/2} \gg 4$ days or $k_{off} \ll 3 \times 10^{-6}$ s^{-1}) or a higher affinity of wild-type NC1 for heme than that of apomyoglobin. By comparison, the k_{off} for sperm whale myoglobin is 8.4×10^{-7} s^{-1}, that for human hemoglobin is 7.1×10^{-6} s^{-1} (26, 27), and the k_{off} for the single-cysteine-ligated heme-sensor proteins such as heme-regulated inhibitor and neuronal PAS protein 2 (NPAS2) are typically $1–6 \times 10^{-3}$ s^{-1} (21). The heme-sensor proteins are a group of heme-binding proteins in which the activity is regulated either directly by heme or by signaling gas molecules such as NO and CO. The slow dissociation of heme from NC1 argues against a model in which activity of DGCR8 is reversibly regulated by its association with heme. Furthermore, the P351A mutation increases the dissociation rate of the heme from NC1, possibly by altering the conformation of the loop containing Cys-352.

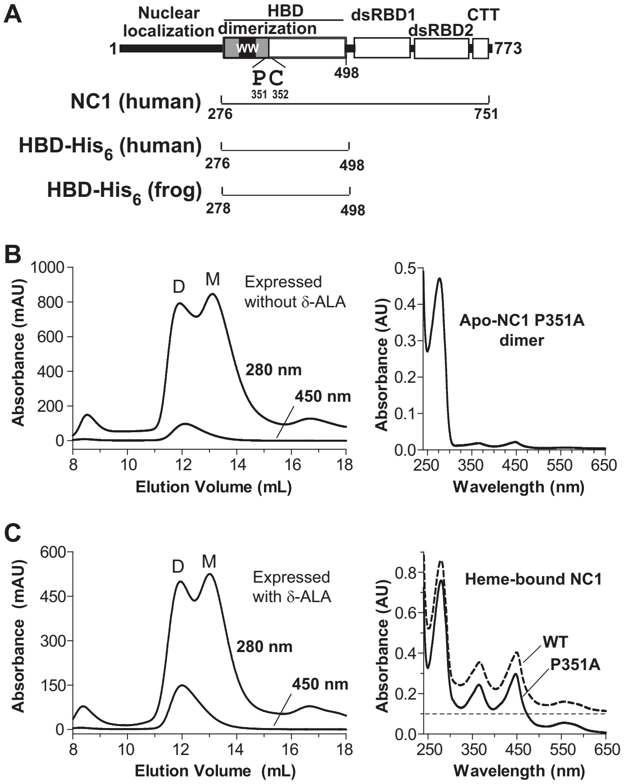

FIGURE 1. **Characterization of the NC1 P351A mutant.** *A*, domain structure of human DGCR8. The double-stranded RNA-binding domains (*dsRBD*) and C-terminal tail (*CTT*) are required and sufficient for cooperative association with pri-miRNAs and for triggering cleavage by the Drosha nuclease. The human NC1, HBD-His$_6$ and frog HBD-His$_6$ protein constructs used in this study are represented by the brackets. *B*, size exclusion chromatogram of NC1 P351A overexpressed in *E. coli*. The 367-nm absorption trace overlaps with the 450-nm one and thus is omitted from the chromatogram drawing. *D* and *M* indicate the dimer and monomer peaks. An electron absorption spectrum of the dimer peak is shown (*right*). *C*, same as *B*, except that 1 mM δ-ALA was added to the bacterial culture at induction.

The P351A-Heme Complex Can Be Reconstituted Using Ferric Heme—It has not been possible to reconstitute the DGCR8-heme complex by incubating heme with the NC1 monomer, the only heme-free form of the wild-type protein currently available. We wondered whether the heme-free P351A dimer would bind heme *in vitro*. We titrated ferric heme into the heme-free P351A dimer and found that they readily formed a specific complex. The P351A-heme complex displayed absorption peaks at 365, 447, and 556 nm (Fig. 3*A*), nearly identical in positions and relative intensities to those of the heme-bound P351A dimer purified from *E. coli*. As the total ferric heme concentration increased, the 447-nm and 556-nm absorption plateaued after ~4.5 μM (Fig. 3, *A* and *B*), close to the P351A dimer concentration of 4.35 μM. This result suggests that each P351A dimer binds one ferric heme molecule. The same protein-heme stoichiometry was previously observed for the wild-type NC1 dimer when it was characterized using HPLC and mass spectrometry (10). In contrast, the absorbance at 365 nm continued to increase linearly after the specific heme-binding site of P351A had been saturated (Fig. 3*B*). Both specific and nonspe-

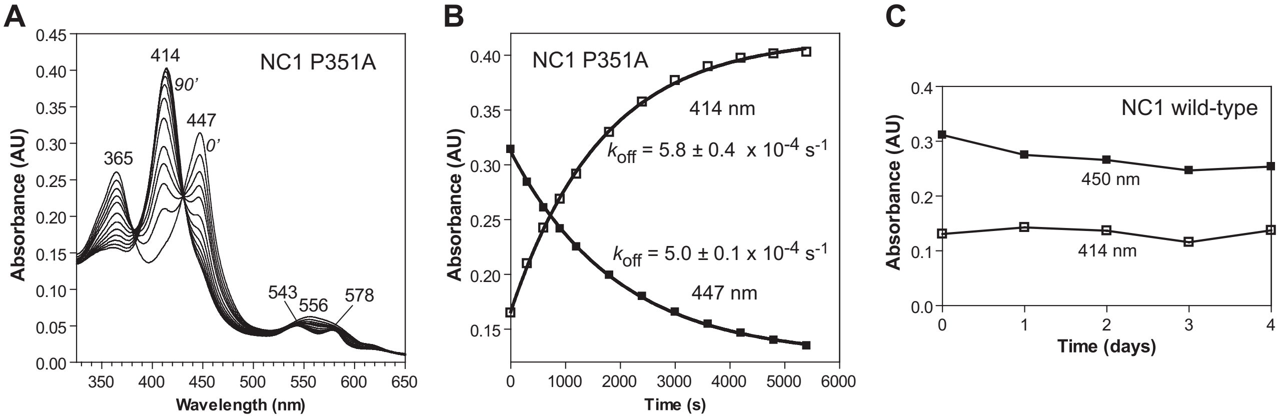

FIGURE 2. **Heme dissociates from the wild-type NC1 slowly, and P351A increases the rate.** Heme-bound NC1, either P351A (*A* and *B*) or wild-type (*C*), was incubated with a 6-fold excess of apomyoglobin at 25 °C. *A*, electronic absorption spectra obtained at 0, 5, 10, 15, 20, 30, 40, 50, 60, 70, 80, and 90 min. *B*, the 414-nm (metmyoglobin) and 447-nm (P351A-heme complex) values shown in *A* were fit to a single exponential function. The rates shown are mean ± S.D. of three independent repeats. *C*, wild-type NC1 showed no significant changes of absorbance at 414 nm and 450 nm at room temperature over 4 days.

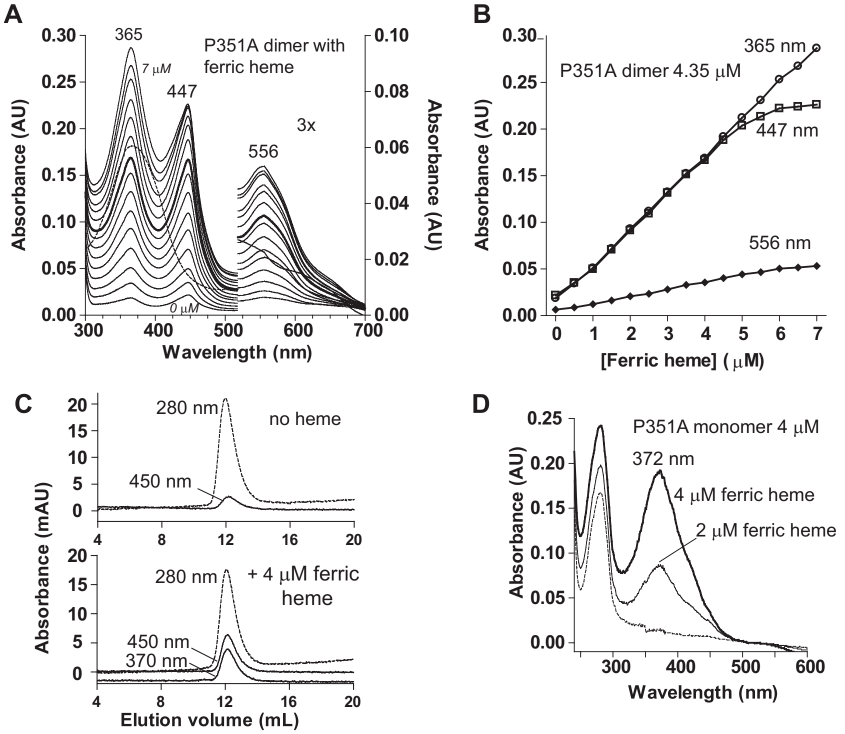

FIGURE 3. **The P351A-heme complex can be reconstituted using ferric heme.** *A*, electronic absorption spectra of titration of ferric heme to heme-free P351A dimer (4.35 μM) in 0.5 μM intervals. The spectrum of ferric heme (5.0 μM) free from proteins is shown using a *dashed line*. *B*, the absorbance values at peak wavelengths are plotted over the concentration of ferric heme added. *C*, SEC of 4 μM heme-free P351A dimer incubated without (*top panel*) or with (*bottom panel*) equal molar of ferric heme. *D*, the P351A monomer shows neither 447- nor 556-nm peaks when incubated with ferric heme.

cific NC1-heme interactions appear to contribute to the 365-nm peak (see below). The reconstituted P351A dimer-ferric heme complex was analyzed using SEC. The absorbances at 280, 370 and 450 nm coeluted in a peak with an elution volume similar to those of the heme-bound and heme-free P351A dimers, indicating that the complex was stable and remained as a dimer (Fig. 3*C*). In contrast to the dimer, the heme-free P351A monomer is incapable of binding ferric heme. Incubation of the heme-free P351A monomer with ferric heme yielded an absorption peak at 372 nm and no Soret peak at ~450 nm (Fig. 3*D*), indicative of nonspecific heme binding. SEC of the heme-free P351A monomer incubated with ferric heme showed that the protein did not coelute with heme (supplemental Fig. S1). Thus, the heme-free P351A monomer cannot bind heme and convert to the heme-bound dimer under the conditions tested. These results confirm the previous finding that dimerization of DGCR8 is required for association with heme (10). Importantly, the *in vitro* heme-binding assays of the P351A dimer were performed in the absence of CO (Fig. 3*A*). The fact that DGCR8 binds ferric heme without CO to give a Soret peak at ~450 nm clearly demonstrates that it must bind heme through a set of interactions distinct from those in ferrous cytochrome P450 bound by CO.

The Hyperporphyrin Spectrum of DGCR8 Matches Those of Ferric Heme with Two Thiolate Ligands—How does DGCR8 interact with ferric heme to generate the 450-nm absorption peak? The similarly intense absorption peaks at ~450 nm and ~366 nm in the DGCR8-heme complexes constitute a hyperporphyrin (split Soret) spectrum. Early studies showed that having a ferrous heme ligated with a cysteine (thiolate) and CO is not the only way to obtain a hyperporphyrin spectrum. Ferric heme in complex with two mercaptide ligands, used as model compounds for hemoproteins, can produce a hyperporphyrin spectrum with the split Soret bands of approximately equal intensity (28). Similarly, ligation of an exogenous thiolate or phosphine compound to ferric heme-thiolate proteins, such as cytochrome P450$_{\mathrm{CAM}}$ and chloroperoxidase, also generates hyperporphyrin spectra with two equally intense Soret bands (supplemental Table S1) (29–31). The electronic absorption spectrum of the NC1-heme complex bears a striking similarity to that of chloroperoxidase with an exogenous thiolate ligand (Fig. 4). Additionally, these studies on cytochrome P450$_{\mathrm{CAM}}$ and chloroperoxidase showed a broad α/β band with its maximum at ~560 nm, similar to the 556-nm peak in the electronic absorption spectrum of the NC1-heme complex (29–31). The match of these electronic absorption spectroscopic signatures suggests that DGCR8 likely binds ferric heme using a cysteine thiolate as the fifth ligand and a sixth ligand that could be either a thiolate or a phosphine. Unlike cytochrome P450 and chloroperoxidase, DGCR8 (both NC1 and HBD-His$_6$) can be purified and stored in the absence of any exogenous thiolate- or phosphine-containing compounds and still maintain its hyperporphyrin spectrum, suggesting that DGCR8 mostly likely uses two endogenous ligands. These ligands could be the side chain of a cysteine (thiolate or thiol) or a methionine (an electron donor similar to the phosphine).

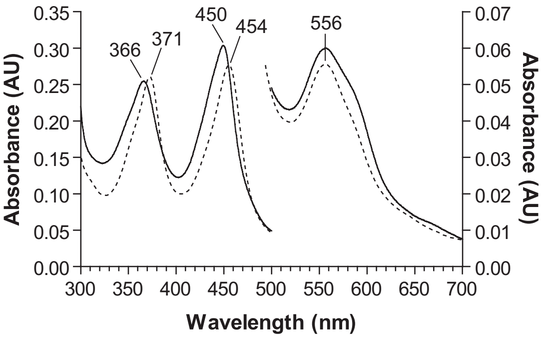

FIGURE 4. **The hyperporphyrin spectrum of DGCR8 likely results from double-cysteine ligation to ferric heme.** The electronic absorption spectrum of ferric heme-bound NC1 dimer (*solid line*) is strikingly similar to that of the ferric chloroperoxidase complexed with ethyl-2-mercaptoacetate (*dashed line*). The latter spectrum was reproduced from an earlier work (31).

MCD and EPR Spectra Reveal Spin, Oxidation, and Coordination States of the Heme in DGCR8—We performed MCD and EPR spectroscopies to further interrogate the chemical environment of the heme in DGCR8. Because these techniques require large amounts of protein, we chose to use the C-terminal His$_6$-tagged heme-binding domain (HBD-His$_6$) from frog DGCR8 (Fig. 1*A*), which can be expressed and purified with a yield of ~1 mg per liter of bacterial culture. The HBD sequences of human and frog DGCR8 are 80% identical. The electronic absorption spectrum of the frog DGCR8 HBD-His$_6$ (Fig. 5, *upper panel*) is nearly identical to those of the human DGCR8 NC1 and HBD-His$_6$ (19) proteins, indicating a close similarity between the chemical environments of their heme cofactors.

The MCD spectrum of the frog DGCR8 HBD-His$_6$ protein, recorded at 4 K, is dominated by two features with peak-crossover-trough positions of 344-368-388 nm and 430-435-448 nm, respectively (Fig. 5, *lower panel*). These features correspond to the blue-shifted and red-shifted split Soret bands in the absorption spectrum. Weaker features are observed in the MCD spectrum between 500 and 750 nm that correspond to the α/β (557 nm) and ligand-to-metal charge-transfer transition (660 nm) peaks in the absorption spectrum. The MCD spectrum of the frog HBD-His$_6$ bears the greatest resemblance to MCD spectra of ferric chloroperoxidase and ferric P450$_{\mathrm{CAM}}$ to which either methanethiol or bis(hydroxymethyl)methylphosphine has been added (5, 31, 32), further supporting the model in which the heme iron in DGCR8 is six-coordinate with one cysteine thiolate ligand, and a sixth ligand that may be either another cysteine (equivalent to the methanethiol) or a methionine (similar to the phosphine). To probe the electronic state of the metal iron, MCD spectra of the frog HBD-His$_6$ were recorded over a range of temperatures (4–50 K). Notably, all of the MCD signals are temperature-dependent, indicating that they originate from a paramagnetic metal. MCD signals at the most intense feature (448 nm) were monitored over a range of magnetic field intensities at several temperatures (2.5–25 K). The resulting magnetic saturation curves at the different temperatures overlap each other (supplemental Fig. S2), indicating that the spin state of the iron is $S = \frac{1}{2}$. This result implies that the heme in DGCR8 contains low-spin Fe(III), with only one unpaired electron.

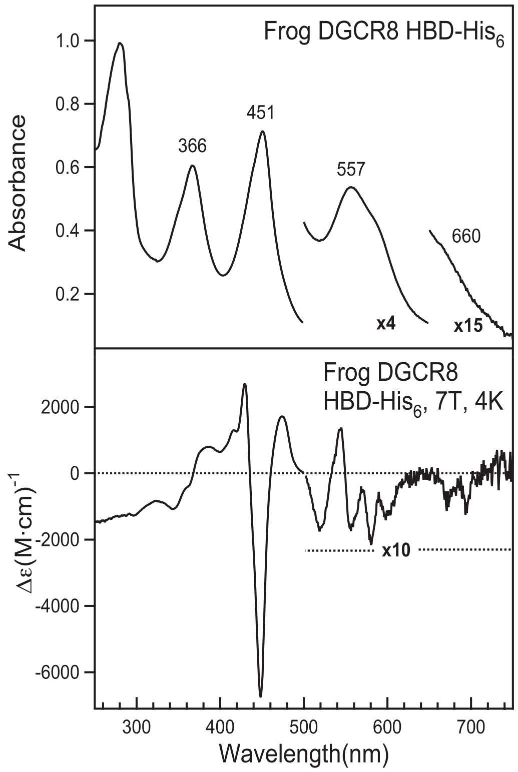

FIGURE 5. **Electronic absorption and MCD spectra of the ferric heme-bound frog DGCR8 HBD-His$_6$.** The electronic absorption spectrum (*top panel*) was taken at room temperature using the dimeric frog HBD-His$_6$ at 12.2 μM concentration in a buffer containing 50 mM EPPS (pH 8.0) and 400 mM NaCl. As a control, an electronic absorption spectrum was also recorded at 4 K and displayed no differences in peak positions and intensity (data not shown). The MCD spectrum (*bottom panel*) was recorded using a protein sample at 24.5 μM in 22.5 mM EPPS (pH 8.0), 180 mM NaCl, and 55% (v/v) glycerol.

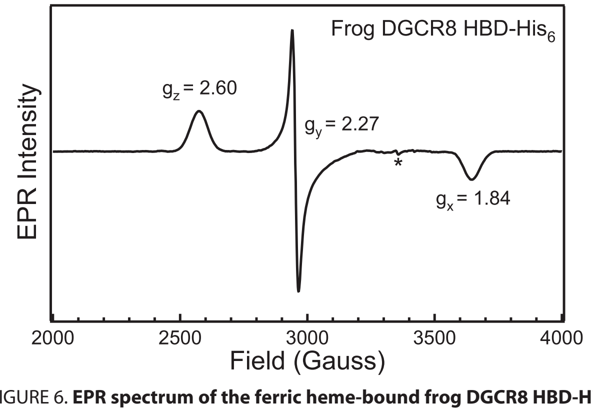

FIGURE 6. **EPR spectrum of the ferric heme-bound frog DGCR8 HBD-His$_6$ protein.** The frog HBD-His$_6$ protein (153 μM) was in 50 mM EPPS (pH 8.0) and 400 mM NaCl. The spectrum represents an average of 10 scans taken at 10 K, with 9.383 GHz microwave frequency, 8.000 G modulation amplitude, 100 kHz modulation frequency, 60 dB receiver gain, 163.84 ms time constant, and a power of 1.002 milliwatt. The *asterisk* represents a signal present in the cavity.

The EPR spectrum of frog DGCR8 HBD-His$_6$ further supports the conclusion that the heme iron is low-spin Fe(III). When the spectrum was scanned from 0–10,000 G, the only signal observed was a narrow rhombic signal centered at ~2900 G (data not shown). A narrow scan (2000–4000 G) around this

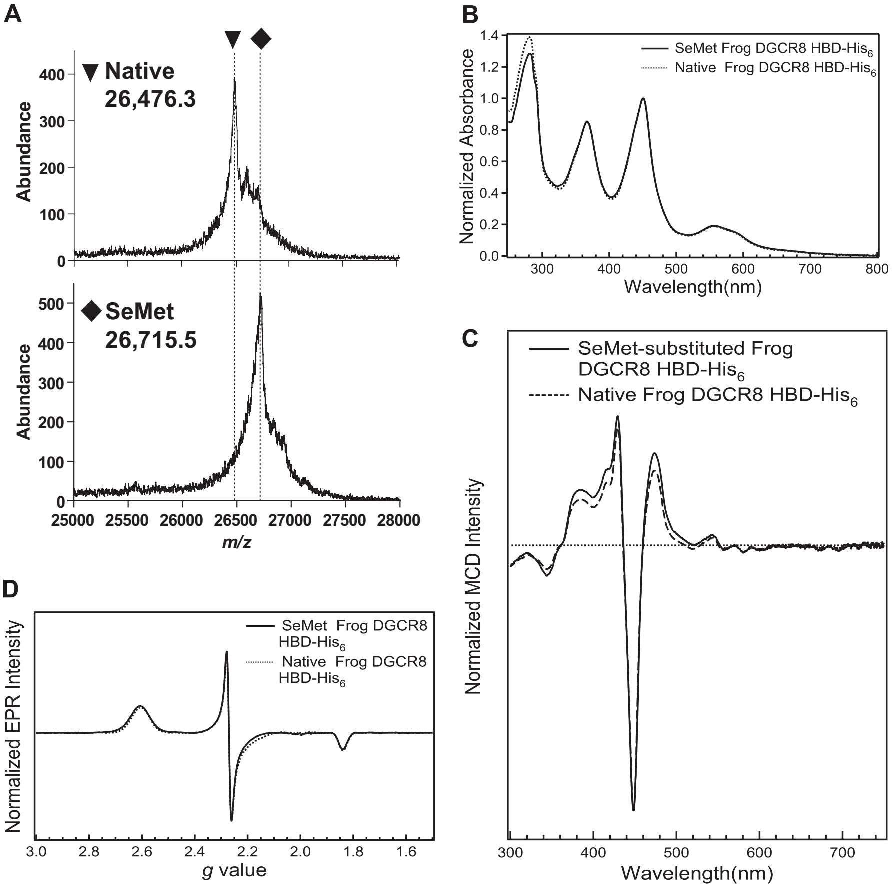

FIGURE 7. **The sixth ligand to ferric heme in DGCR8 is not a methionine.** *A*, mass spectra of native and SeMet-labeled frog DGCR8 HBD-His$_6$ indicates nearly complete substitution of Met by SeMet. The difference in their *m/z* peaks, 239, is consistent with all five methionine residues substituted with SeMet. *B*, the normalized electronic absorption spectrum of the SeMet-substituted frog HBD-His$_6$ is nearly identical to that of the native protein (Fig. 5, *top panel*). *C*, the normalized MCD spectrum of the SeMet-substituted frog HBD-His$_6$ is nearly identical to that of the native protein (Fig. 5, *bottom panel*). *D*, comparison of the EPR spectra of the native and SeMet-labeled frog HBD-His$_6$. The spectral parameters for the SeMet-labeled frog HBD-His$_6$ were the same as those of the native listed in Fig. 6, except that the ferric heme concentration of the sample was 142 μM and the spectrum was acquired with a 9.384 GHz microwave frequency.

rhombic signal gave the three *g* values: $g_z = 2.60$, $g_y = 2.27$, and $g_x = 1.84$ (Fig. 6). The presence of a single rhombic signal in the $g = 2$ region reveals that the iron in DGCR8 is clearly low spin Fe(III). Analysis of the calculated crystal field parameters of tetragonality (Δ/λ, degree of electronic donation at the iron center) and rhombicity (V/Δ, degree of geometric distortion of the iron center) places the frog Fe(III) HBD-His$_6$ in the same region as native ferric chloroperoxidase in the Blumberg-Peisach "truth" diagrams (33), in closest proximity to Fe(III) chloroperoxidase thiol adducts such as thioacetic acid and β-mercaptoethanol and the Fe(III) chloroperoxidase bis(hydroxymethyl)methyl phosphine adduct (32, 34) (supplemental Table S2 and Fig. S3). Although a number of thiol adducts of ferric chloroperoxidase and ferric P450$_{\mathrm{CAM}}$ display multiple overlapping, rhombic EPR signals, the frog Fe(III) HBD-His$_6$ only displays a single rhombic signal, suggesting a well defined, homogenous environment about the DGCR8 heme.

The Sixth Axial Ligand of Heme in DGCR8 Is Not a Methionine—Because both thiols and phosphines ligated to ferric chloroperoxidase and P450 generate hyperporphyrin spectra (5, 32), we considered the possibility that one of the ligands coordinated to the heme in DGCR8 is the thioether of a methionine side chain. We examined the spectroscopic properties of the frog HBD-His$_6$ labeled with SeMet, with the expectation that if methionine were a ligand to the heme iron, its substitution with the selenium analog would be evident in the heme spectra. A MALDI-TOF mass spectrum of the SeMet-labeled HBD-His$_6$ showed a single major peak with a *m/z* value greater than that of the unlabeled protein by 239, within experimental error ($\leq \pm 13$) of the expected increase of 234 when all five methionine residues in the frog HBD-His$_6$ sequence are substituted with SeMet (Fig. 7*A*). This result indicated that nearly complete substitution of methionine with SeMet was achieved. Importantly, the electronic absorption spectrum of SeMet-labeled frog HBD-His$_6$ shows no changes in peak positions relative to the native construct (Fig. 7*B*). In contrast, previous studies showed that substitution of a methionine ligand with SeMet caused the Soret α and β peaks to red-shift on the range of 3–7 nm (35, 36). The signals in the MCD and EPR spectra of SeMet-labeled HBD-His$_6$ were also unchanged from those of the native Met-containing protein (Fig. 7, *C* and *D*). The large difference in spin-orbit coupling between Se and S is expected to give rise to significantly different spectral features; that the MCD and

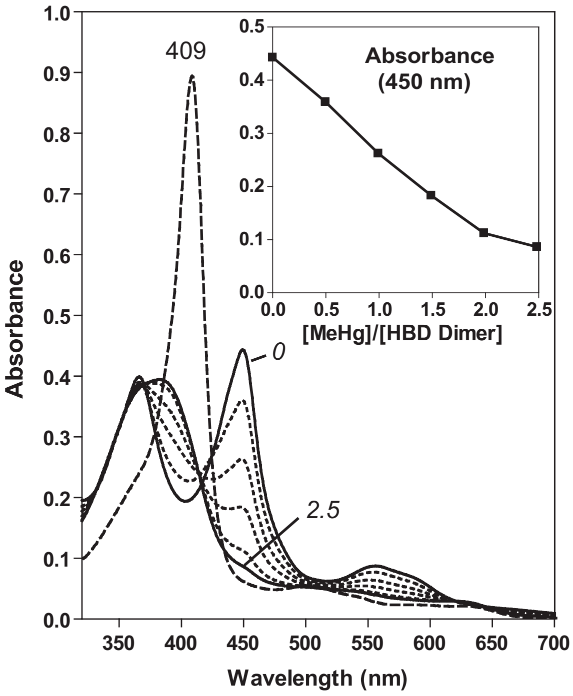

FIGURE 8. **Cys-352 from both subunits in a DGCR8 dimer serve as the axial ligands to ferric heme.** Electronic absorption spectra of $hHBD^{C430S}$ (10 μM) titrated with MeHgAc. The MeHgAc was added at steps of 0.5 molar equivalent of ferric heme-bound HBD dimer. The spectra of the starting point (0 molar equivalent of MeHgAc) and end point (2.5 molar equivalents) are shown in solid traces and the intermediate steps in *dotted lines*. After the MeHgAc titration, 1.2 molar equivalents of apomyoglobin (relative to the HBD dimer) were added, and the electronic absorption spectrum was immediately recorded (*dashed line*). The *inset* shows absorbance at 450 nm over the course of the titration.

EPR spectra of the two proteins are the same argues that methionine is not the sixth ligand to the heme in DGCR8. Thus, DGCR8 must use two endogenous cysteine residues to ligate heme.

Cys-352 from both Subunits in a DGCR8 Dimer Serve as the Axial Ligand of Heme—Which cysteine residues does DGCR8 use for ligating heme? There are seven cysteine residues in the NC1 (human) sequence; however, only two of them, Cys-352 and Cys-430, are located in the HBD region. Available data suggest that the HBD binds heme in the same way as NC1 (19) and that Cys-430 is not required for association with heme (10). Thus, Cys-352 is the only cysteine residue in human DGCR8 that can serve as a ligand for heme. Furthermore, heme binds to dimeric DGCR8, suggesting a model in which the two Cys-352 from both subunits serve as the axial ligands of heme.

We directly tested this model using a mercury titration strategy. Mercury compounds bind thiol groups with high affinities and thus can disrupt the interaction between cysteine ligands and the heme cofactor. To simplify the interpretation, we first mutated Cys-430 to serine in the context of human DGCR8 HBD-His_6 ($hHBD^{C430S}$) so that Cys-352 was the only cysteine present in the sequence. Not surprisingly, the mutant exhibited an electronic absorption spectrum with peak positions and relative intensities indistinguishable from those of the wild-type protein (Fig. 8). This result confirms experimentally that Cys-352 is the only cysteine residue that can ligate heme. Titration of the $hHBD^{C430S}$ dimer with MeHgAc led to a linear decrease of the peaks at 366, 450, and 556 nm until 2.0 molar equivalents (relative to $hHBD^{C430S}$ dimer) was reached. The peaks disappeared completely after addition of 2.5 molar equivalents of MeHgAc (Fig. 8). Concomitantly, there appeared a broad electronic absorption peak centered at 385 nm resembling that of free ferric heme (Fig. 8), suggesting that the heme was released from $hHBD^{C430S}$ after both Cys-352 ligands dissociated from the heme iron. Consistent with this interpretation, when 1.2 molar equivalents of apomyoglobin were added to the mercury-treated $hHBD^{C430S}$ dimer, the Soret peak shifted to 409 nm, characteristic of ferric myoglobin (metmyoglobin). Because each methylmercury generally binds one cysteine in a protein, these results indicate that both Cys-352 residues in the $hHBD^{C430S}$ dimer are important for association with heme.

Heme Binds to DGCR8 Expressed in Eukaryotic Cells—We have studied the DGCR8-heme interaction using overexpression in *E. coli* or biochemical reconstitution; however, the question remains: does this family of proteins, found only in animals, bind heme in eukaryotic cells? Because overexpression of DGCR8 in mammalian cells is subjected to multiple levels of regulation (37) and appears to be cytotoxic, we addressed this question by expressing NC1 with a N-terminal His_6 tag (His_6-NC1) in Sf9 insect cells (Fig. 9). The purified His_6-NC1 protein, which appeared largely as a single band, displayed absorption peaks at 366, 450, and 556 nm (Fig. 9*A*), identical to those of the NC1-ferric heme complex expressed in bacteria. This His_6-NC1 protein sample was further analyzed using SEC (Fig. 9*B*). The chromatogram showed that the insect cell-expressed His_6-NC1 eluted at a volume expected for the NC1 dimer, and no monomer peak was visible, suggesting that the His_6-NC1 existed nearly exclusively as a dimer. SEC also allowed the small amounts of protein and nucleic acid impurities to be removed from His_6-NC1. Based on the absorbance ratio A_{450}/A_{280}, the heme occupancy of the insect cell-expressed His_6-NC1 was estimated to be ~50%. We tested the His_6-NC1 using reconstituted pri-miRNA processing assays. This assay revealed that the activity of the Sf9-expressed His_6-NC1 was comparable with that of the heme-bound NC1 dimer expressed in bacteria (Fig. 9*C*). Given these data, we conclude that active DGCR8 proteins expressed in eukaryotic cells can bind heme.

DISCUSSION

Here we present biochemical and spectroscopic characteristics of the DGCR8-heme complex, revealing interactions distinct from those of other known cysteine-ligated hemoproteins. The DGCR8-heme complex displays an absorption spectrum with an intense peak near 450 nm in the absence of CO. Our data reveal that the heme iron is in the ferric state and that DGCR8 ligates to the heme iron using two endogenous cysteine side chains (Fig. 10).

Our recent crystal structure of the N-terminal dimerization subdomain of the human DGCR8 HBD revealed a symmetric dimer with an extensive dimerization interface mediated mostly by hydrophobic interactions (19). Although our structure does not contain heme, it shows that the two Cys-352 residues in the dimer are within van der Waals contacting distance

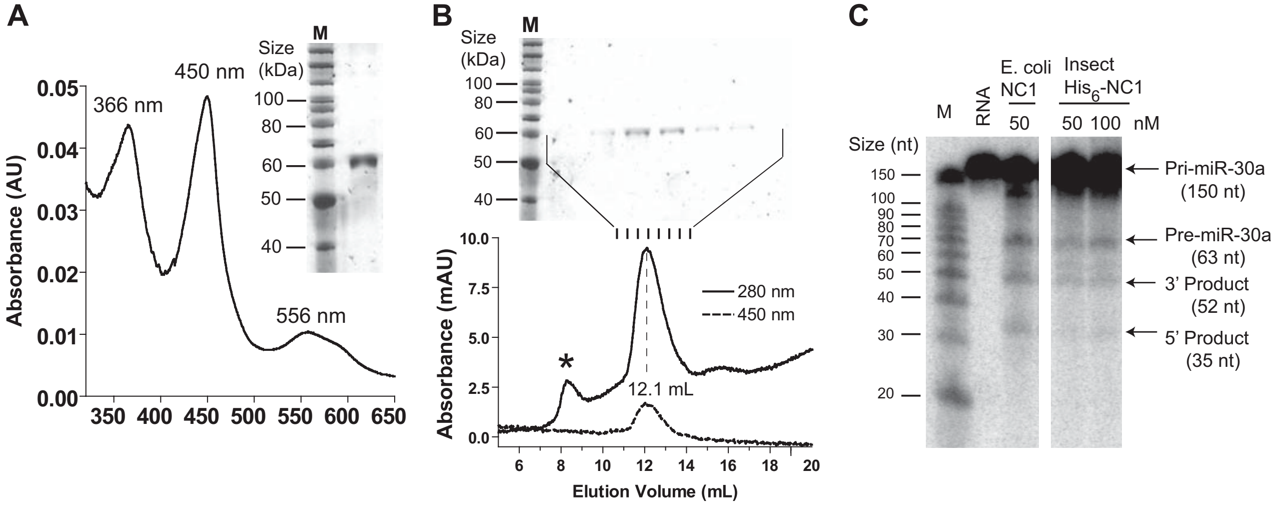

FIGURE 9. **DGCR8 binds heme in insect cells.** *A*, electronic absorption spectrum of His$_6$-NC1 (human) expressed in Sf9 cells using baculovirus. His$_6$-NC1 was purified using Ni affinity chromatography followed by cation exchange chromatography. The image of a silver-stained SDS-polyacrylamide gel is shown. *B*, analytical size exclusion chromatogram of the His$_6$-NC1 sample used in *A*, with the silver-stained gel of the SEC fractions shown above. The peak indicated by the asterisk contains mainly nucleic acids. *C*, His$_6$-NC1 expressed in insect cells was used in reconstituted pri-miRNA processing assays. A pri-miR-30a fragment was uniformly labeled with α-^{32}P-UTP and was incubated with recombinant Drosha and NC1 at indicated concentrations. The reactions were analyzed using denaturing gel electrophoresis and autoradiography.

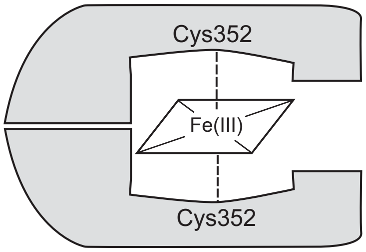

FIGURE 10. **Schematic of how the DGCR8 HBD binds ferric heme.**

of each other. A conformational change of the loop containing Cys-352 likely occurs upon association of the protein with heme and should be able to position the two Cys-352 side chains on opposite sides of the heme plane for axial ligation. Here we show that the DGCR8-heme complexes are remarkably stable (Fig. 2*C*). Proper positioning of the two Cys-352 side chains within a robust dimer likely contributes to this stability.

This study establishes DGCR8 as the first example of a heme protein with two endogenous cysteines, Cys-352 from both subunits, as axial ligands. Our investigation revealed that Pro-351 is important for the stability of the DGCR8-heme complex. Pro-351 and Cys-352 together are similar to the heme regulatory motif, which contains a Cys-Pro sequence but in the opposite order. Heme regulatory motifs directly bind heme and are found in repeats in a range of heme-regulated proteins, such as the yeast transcription activator HAP1 and the human δ-aminolevulinic acid synthase, heme oxygenase-2, heme-regulated eukaryotic initiation factor 2α kinase, and Rev-erbβ (38–42). It is possible that in both motifs the proline helps to position the cysteine side chain in an appropriate geometry for ligation and for either catalytic or regulatory functions. However, unlike the heme regulatory motifs, the Pro-351 and Cys-352 residues are not sufficient for DGCR8 to associate with heme, because dimerization and Trp-329 in the WW motif are also required for heme binding.

Our previous (10) and current studies clearly demonstrated that dimerization is required for association with heme. The heme-free forms of NC1 expressed in *E. coli* are either mostly monomeric (for the wild type and the C351A mutant) (10) or a mixture of monomers and dimers (for P351A, this study). We have never observed monomeric NC1 to bind heme and dimerize *in vitro*, regardless of whether it is the wild type or the P351A mutant, (Fig. 3*D* and data not shown). Furthermore, it appears that the monomeric species of NC1 is only generated in *E. coli* and not in insect cells (Fig. 9*B*). A careful inspection of the SDS-12% polyacrylamide gels of the bacteria-expressed NC1 revealed that the monomer contained low-molecular-weight protein species comigrating with the dye bromphenol blue (supplemental Fig. S4, *A* and *B*). On a SDS-15% polyacrylamide gel, these protein species appeared as an approximately 10-kDa band with a smear underneath (supplemental Fig. S4*C*), indicating that they are heterogeneous. Furthermore, our crystal structure of the N-terminal sub-domain of human DGCR8 HBD shows an extensive dimerization interface that does not appear to be easily disrupted in non-denaturing conditions. Thus, the "monomeric" NC1 may partially, if not entirely, consist of heterodimers in which one subunit has been non-specifically cleaved in the central loop region (19) of its HBD, in the *E. coli* cells or during purification, into a fragment that retains the dimerization domain. It is possible that in human cells both heme-free and heme-bound DGCR8 are dimers.

The current study reveals important features of the DGCR8-heme interaction that will help us to understand its potential physiological function in regulating miRNA maturation. We have two working models. In the first model, the DGCR8-heme complex may serve as a sensor for cellular or environmental signals (the heme-based sensor model) (43). In a heme-based sensor, heme is stably bound to a protein partner, which is consistent with the slow dissociation of heme from NC1 observed here (Fig. 2*C*). In the second, heme itself may serve as the ligand to activate miRNA processing (the heme sensing model) (21). The slow dissociation of heme from NC1 argues against a heme sensing model in which heme reversibly binds to DGCR8 and regulates its activity. However, it is still possible that heme may bind and activate DGCR8 while the signal is turned off through protein degradation.

REFERENCES

1. Perutz, M. F., Wilkinson, A. J., Paoli, M., and Dodson, G. G. (1998) *Annu. Rev. Biophys. Biomol. Struct.* **27,** 1–34
2. Derbyshire, E. R., and Marletta, M. A. (2009) *Handb. Exp. Pharmacol.* **191,** 17–31
3. Ortiz de Montellano, P. R. (2000) *Curr. Opin. Chem. Biol.* **4,** 221–227
4. Ortiz de Montellano, P. R. (2004) *Cytochrome P450: Structure, Mechanism, and Biochemistry*, 3rd Ed., Springer, New York
5. Dawson, J. H., and Sono, M. (1987) *Chem. Rev.* **87,** 1255–1276
6. Rousseau, D. L., Li, D., Couture, M., and Yeh, S. R. (2005) *J. Inorg. Biochem.* **99,** 306–323
7. Udit, A. K., and Gray, H. B. (2005) *Biochem. Biophys. Res. Commun.* **338,** 470–476
8. Dawson, J. H. (1988) *Science* **240,** 433–439
9. Paoli, M., Marles-Wright, J., and Smith, A. (2002) *DNA Cell. Biol.* **21,** 271–280
10. Faller, M., Matsunaga, M., Yin, S., Loo, J. A., and Guo, F. (2007) *Nat. Struct. Mol. Biol.* **14,** 23–29
11. Denli, A. M., Tops, B. B., Plasterk, R. H., Ketting, R. F., and Hannon, G. J. (2004) *Nature* **432,** 231–235
12. Gregory, R. I., Yan, K. P., Amuthan, G., Chendrimada, T., Doratotaj, B., Cooch, N., and Shiekhattar, R. (2004) *Nature* **432,** 235–240
13. Han, J., Lee, Y., Yeom, K. H., Kim, Y. K., Jin, H., and Kim, V. N. (2004) *Genes Dev.* **18,** 3016–3027
14. Landthaler, M., Yalcin, A., and Tuschl, T. (2004) *Curr. Biol.* **14,** 2162–2167
15. Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Rådmark, O., Kim, S., and Kim, V. N. (2003) *Nature* **425,** 415–419
16. Kim, V. N., Han, J., and Siomi, M. C. (2009) *Nat. Rev. Mol. Cell Biol.* **10,** 126–139
17. Faller, M., and Guo, F. (2008) *Biochim. Biophys. Acta* **1779,** 663–667
18. Faller, M., Toso, D., Matsunaga, M., Atanasov, I., Senturia, R., Chen, Y., Zhou, Z. H., and Guo, F. (2010) *RNA* **16,** 1570–1583
19. Senturia, R., Faller, M., Yin, S., Loo, J. A., Cascio, D., Sawaya, M. R., Hwang, D., Clubb, R. T., and Guo, F. (2010) *Protein Sci.* **19,** 1354–1365
20. Omura, T. (2005) *Biochem. Biophys. Res. Commun.* **338,** 404–409
21. Igarashi, J., Kitanishi, K., Martinkova, M., Murase, M., Iizuka, A., and Shimizu, T. (2008) *Acta Chim. Slov.* **55,** 67–74
22. Stern, J. O., and Peisach, J. (1974) *J. Biol. Chem.* **249,** 7495–7498
23. Van Duyne, G. D., Standaert, R. F., Karplus, P. A., Schreiber, S. L., and Clardy, J. (1993) *J. Mol. Biol.* **229,** 105–124
24. Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D., and Bairoch, A. (2003) *Nucleic Acids Res.* **31,** 3784–3788
25. Hargrove, M. S., Krzywda, S., Wilkinson, A. J., Dou, Y., Ikeda-Saito, M., and Olson, J. S. (1994) *Biochemistry* **33,** 11767–11775
26. Hargrove, M. S., Barrick, D., and Olson, J. S. (1996) *Biochemistry* **35,** 11293–11299
27. Hargrove, M. S., Singleton, E. W., Quillin, M. L., Ortiz, L. A., Phillips, G. N., Jr., Olson, J. S., and Mathews, A. J. (1994) *J. Biol. Chem.* **269,** 4207–4214
28. Ruf, H. H., and Wende, P. (1977) *J. Am. Chem. Soc.* **99,** 5499–5500
29. Yu, C., Gunsalus, I. C., Katagiri, M., Suhara, K., and Takemori, S. (1974) *J. Biol. Chem.* **249,** 94–101
30. Sono, M., Andersson, L. A., and Dawson, J. H. (1982) *J. Biol. Chem.* **257,** 8308–8320
31. Sono, M., Dawson, J. H., and Hager, L. P. (1984) *J. Biol. Chem.* **259,** 13209–13216
32. Sono, M., Dawson, J. H., and Hager, L. P. (1985) *Inorg. Chem.* **24,** 4339–4343
33. Blumberg, W. E., and Peisach, J. (1971) *Adv. Chem.* **100,** 271–291
34. Sono, M., Hager, L. P., and Dawson, J. H. (1991) *Biochim. Biophys. Acta* **1078,** 351–359
35. Wallace, C. J., and Clark-Lewis, I. (1992) *J. Biol. Chem.* **267,** 3852–3861
36. Low, D. W., Hill, M. G., Carrasco, M. R., Kent, S. B., and Botti, P. (2001) *Proc. Natl. Acad. Sci. U.S.A.* **98,** 6554–6559
37. Han, J., Pedersen, J. S., Kwon, S. C., Belair, C. D., Kim, Y. K., Yeom, K. H., Yang, W. Y., Haussler, D., Blelloch, R., and Kim, V. N. (2009) *Cell* **136,** 75–84
38. Zhang, L., and Guarente, L. (1995) *EMBO J.* **14,** 313–320
39. McCoubrey, W. K., Jr., Huang, T. J., and Maines, M. D. (1997) *J. Biol. Chem.* **272,** 12568–12574
40. Yi, L., and Ragsdale, S. W. (2007) *J. Biol. Chem.* **282,** 21056–21067
41. Igarashi, J., Murase, M., Iizuka, A., Pichierri, F., Martinkova, M., and Shimizu, T. (2008) *J. Biol. Chem.* **283,** 18782–18791
42. Gupta, N., and Ragsdale, S. W. (2011) *J. Biol. Chem.* **286,** 4392–4403
43. Gilles-Gonzalez, M. A., and Gonzalez, G. (2005) *J. Inorg. Biochem.* **99,** 1–22

Supplemental figures, figure legends and tables

DiGeorge Critical Region 8 (DGCR8) Is a Double-Cysteine-Ligated Heme Protein

Ian Barr[1], Aaron T. Smith[2], Rachel Senturia[1], Yanqiu Chen[1], Brooke D. Scheidemantle[1], Judith N. Burstyn[2,3], Feng Guo[1,3]

From [1] Department of Biological Chemistry, David Geffen School of Medicine, Molecular Biology Institute, University of California, Los Angeles, CA 90095, USA, and

[2] Department of Chemistry, University of Wisconsin–Madison, Madison, Wisconsin 53706, USA

[3] Address correspondence to: Feng Guo, Ph.D., 611 Charles E. Young Dr. E., 202 Boyer Hall, Los Angeles, CA 90095. Fax: 310-206-7286; E-mail: fguo@mbi.ucla.edu. Or Judith Burstyn, 1101 University Avenue, Madison, Wisconsin 53706. Fax: F608-262-6143; E-mail: burstyn@chem.wisc.edu

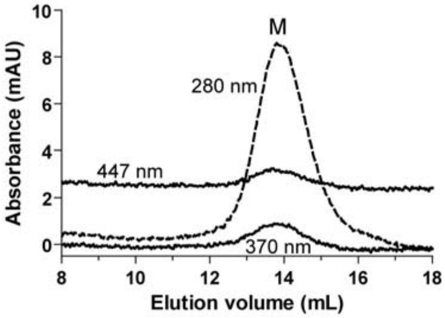

Figure S1. Monomeric P351A did not bind heme.

Size exclusion chromatogram of 4 µM heme-free P351A monomer incubated with equimolar ferric heme. Monomeric heme-free P351A was purified following the procedure as described (1), except with an additional round of size exclusion chromatography to remove the residual dimers. The 447-nm curve was nudged upwards to improve clarity. The monomer peak is labeled as "M". A 370-nm absorbing species co-eluted with the P351A monomer; this species most likely represents non-specifically bound heme, and may constitute 15% of the heme that was added in the incubation.

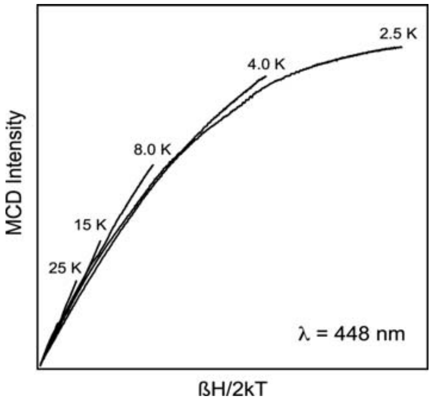

Figure S2. Magnetic saturation curves taken for the frog DGCR8 HBD-His$_6$.

The field dependence of the MCD intensity at 448 nm was recorded at 2.5, 4.0, 8.0, 15 and 25K. The curves were normalized to the most intense data point (2.5K, 7T). The nesting of these curves is consistent with a low spin Fe(III), S=1/2 state for the heme iron.

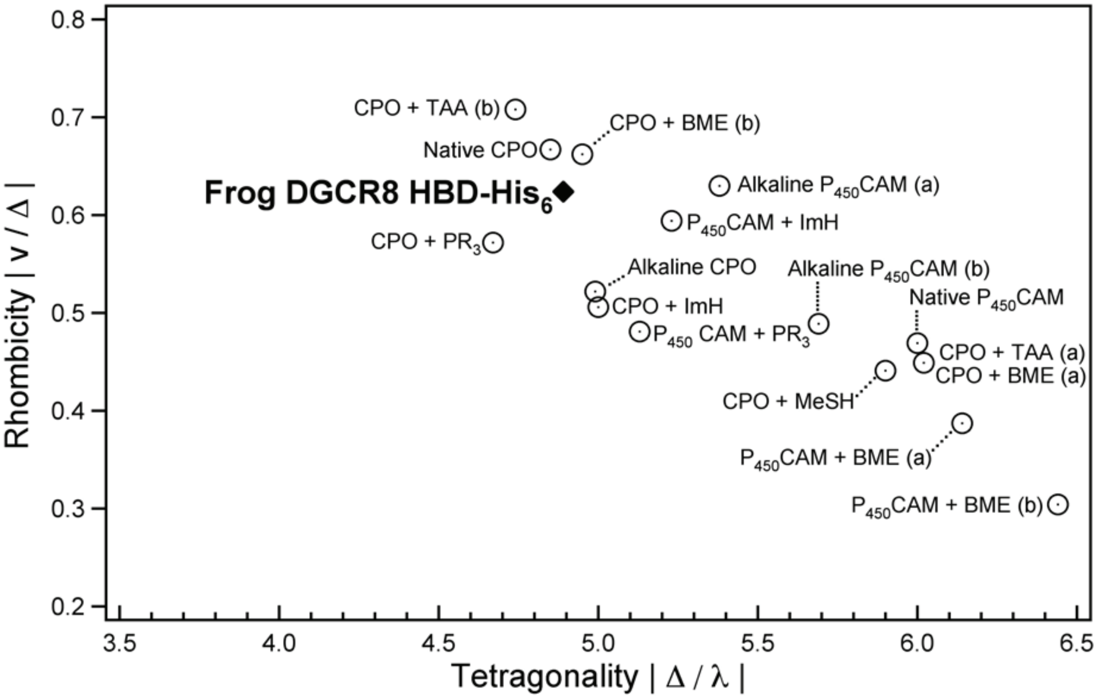

Figure S3. Blumberg-Peisach correlation diagram showing the position of ferric frog DGCR8 HBD-His$_6$ among other sulfur- and phosphorus-donor complexes of ferric chloroperoxidase and cytochrome P450$_{CAM}$.

The plotted values of rhombicity (|v/Δ|) versus tetragonality (|Δ/λ|) show that the environment of the Fe(III) heme in frog DGCR8 HBD-His$_6$ is most similar to that of native chloroperoxidase to which exogenous thiols or phosphines have been added. |v/Δ| and |Δ/λ| were calculated directly from the EPR *g* values presented in reference (2). The following improper axis system was used, as outlined in reference (3): $g_x = -g_y$; $g_y = g_z$; and $g_z = -g_x$. Rhombicity and tetragonality values were each calculated as outlined in (7) after the improper axis transformation was applied: $V/\lambda = (g_x/(g_z + g_y)) + (g_y/(g_z - g_x))$; $\Delta/\lambda = |(g_x/2(g_z + g_y)) + (g_z/(g_y - g_x)) - (g_y/2(g_z - g_x))|$; $V/\Delta = (V/\lambda) / (\Delta/\lambda)$. The following abbreviations are used: CPO: chloroperoxidase; P450$_{CAM}$: cytochrome P450$_{CAM}$; TAA: thioacetic acid; BME: β-mercaptoethanol; ImH: imidazole; MeSH: methanethiol; PR$_3$: bis(hydroxymethyl)-methyphosphine. "(a)" and "(b)" in the diagram refer to different sets of multiple, overlapping rhombic signals that are observed for a single sample.

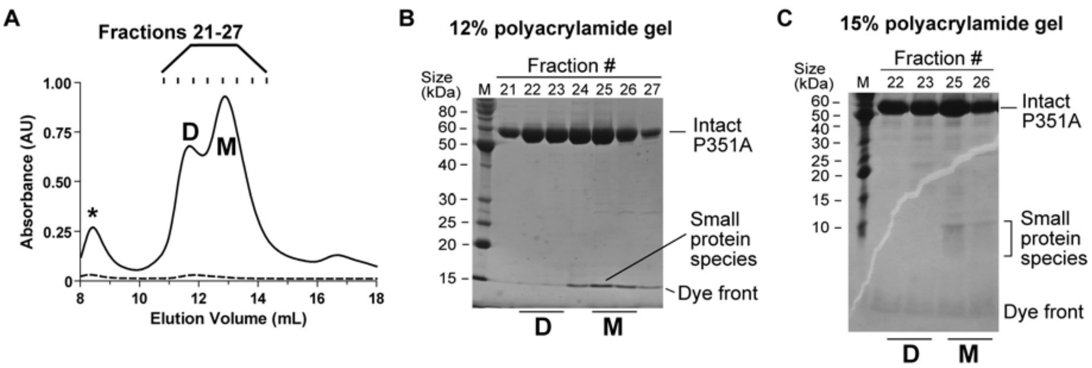

Figure S4. The "monomer" fractions of NC1 P351A expressed in *E. coli* may consist of heterodimers.

The NC1 P351A mutant was expressed in *E. coli* and purified using the procedure described in the main text. (*A*) Size exclusion chromatogram in the last step of the purification procedure. "D" and "M" indicate the peaks assigned to be dimers and monomers, respectively. The asterisk indicates an impurity (mostly nucleic acids) from the bacterial extract. The SDS-12% (*B*) and 15% (*C*) polyacrylamide gels of the chromatographic fractions, stained with Coomassie Brilliant Blue G-250, showed small protein species. These species co-migrated with the dye front on the 12% gel (*B*) and were resolved into a smear with the most intensity centered at around 10 kDa (*C*). Western blotting using an anti-DGCR8 antibody (M.G. and M.F., data not shown) indicated that the small protein species are fragments of DGCR8, thus might have originated from proteolytic cleavage of a NC1 subunit during bacterial expression or purification. Similar observations were made for the wild-type NC1 protein as well.

Supplemental Tables

Table S1. Electronic absorption peaks and spin states of native DGCR8 proteins and of sulfur- and phosphorus-donor complexes of ferric chloroperoxidase and cytochrome $P450_{CAM}$.

		Soret[d]				
Fe(III) proteins and ligands	**pH**	**Hyper blue**	**Hyper red**	**α/β**	**Spin state**	**Ref**
Human DGCR8 NC1 Native	8.0	366	450	556	LS[a]	This work
Frog DGCR8 HBD Native	8.0	366	451	557	LS	This work
Chloroperoxidase						
Native	3-7	399		514/542	LS	(4)
Thiol/thiolate						
H_2S	6-7	369	449	NR[b]	LS	(4)
H_3CSH	6.0	372	455	NR	LS	(4)
$HO(CH_2)_2SH$	6.0	380	455	NR	HS/LS[c]	(4)
$H_3CH_2COC(O)CH_2SH$	6.0	371	454	556	LS	(4)
Phosphine						
$(HOCH_2)_2PCH_3$	6.0, 6.5	376	450	553	LS	(5)
Thioether						
H_3CSCH_3	3-6	417		NR	HS/LS	(4)
Disulfide						
H_3CSSCH_3	3-6	417		NR	HS/LS	(4)
Cytochrome $P450_{CAM}$						
Native (camphor-free)	7.0	417		536/569	LS	(6)
Thiol/thiolate						
H_2S	7.0	390	467	564	LS	(6)
$H_3C(CH_2)_2SH$	9.1	377	466		LS	(6)
$HO(CH_2)_2SH$	8.2	376	464	558	LS	(6)
$ClC(CH)_4CSH$	7.0	381	463	560	LS	(6)
Phosphine						
$(HOCH_2)_2PCH_3$	7.3	375	446	553	LS	(5)
Thioether						
H_3CSCH_3	7.0	424		538/570	LS	(6)
Disulfide						
H_3CSSCH_3	7.0	418		536/568	LS	(6)

[a] LS: low-spin; HS: high-spin. [b] NR: not reported. [c] HS/LS indicates that the protein exists as an admixture of high- and low-spin states. [d] Hyper red and hyper blue refer to the two components of the split Soret. A single value indicates the absence of a hyperporphyrin spectrum.

Table S2. EPR parameters (*g* values) of the native Frog DGCR8 HBD and of sulfur- and phosphorus-donor complexes of ferric chloroperoxidase and cytochrome P450$_{CAM}$.

Fe(III) protein	pH	g_z	g_y	g_x	Ref
Frog DGCR8 HBD Native	8.0	2.60	2.27	1.84	This work
Chloroperoxidase					
Native	6.0	2.62	2.26	1.83	(2)
	9.5	2.54	2.28	1.85	(2)
Thiol/thiolate					
H_3CSH	3.0	2.45	2.27	1.91	(2)
$H_3CH_2COC(O)H_2CSH$	6.0	2.37	2.25	1.94	(2)
		[*2.43*[a]	*2.26*[a]	*1.91*[a]]	(2)
$H_3CC(O)SH$	3.0	2.66	2.26	1.82	(2)
		[*2.44*		*1.91*]	(2)
$HO(CH_2)_2SH$	6.0	2.61	2.26	1.84	(2)
		[*2.44*		*1.91*]	(2)
Phosphine					
$(HOCH_2)_2PCH_3$	7.0	2.59	2.28	1.82	(5)
Thioether					
H_3CSCH_3	NR[b]	2.62	2.26	1.83	(7)
Disulfide					
H_3CSSCH_3	NR	2.62	2.26	1.83	(7)
Cytochrome P450$_{CAM}$					
Native (camphor-free)	5.5-9	2.44	2.25	1.91	(6)
Thiol/thiolate					
$H_3C(CH_2)_2SH$	7.0	2.41	2.24	1.92	(6)
		[*2.34*		*1.94*]	(6)
$HO(CH_2)_2SH$	7.0	2.42	2.27	1.92	(6)
		[*2.38*	*2.27*	*1.93*]	(6)
		[*2.52*	*2.27*	*1.90*]	(6)
$(C_6H_5)CH_2SH$	6-9.5	2.39	2.25	1.94	(6)
		[*2.44*	*2.25*	*1.92*]	(6)
Phosphine					
$(HOCH_2)_2PCH_3$	7.0	2.51	2.28	1.86	(5)
Thioether					
H_3CSCH_3	7.0	2.50	2.27	1.89	(6)
Disulfide					
H_3CSSCH_3	7.0	2.42	2.25	1.92	(6)

[a] Values in brackets represent minor signals present. In most cases, the relationship of the major signal to the minor signal was not explored. [b] NR: not reported.

REFERENCES

1. Faller, M., Matsunaga, M., Yin, S., Loo, J. A., and Guo, F. (2007) *Nat. Struct. Mol. Biol.* **14**, 23-29
2. Sono, M., Hager, L. P., and Dawson, J. H. (1991) *Biochim Biophys Acta* **1078**, 351-359
3. Palmer, G. (1983) in: *Iron Porphyrins, Part II* (Lever, A. B. P., and Gray, H. B. eds) pp. 45-88, VCH Publishers, New York
4. Sono, M., Dawson, J. H., and Hager, L. P. (1984) *J. Biol. Chem.* **259**, 13209-13216
5. Sono, M., Dawson, J. H., and Hager, L. P. (1985) *Inorg. Chem.* **24**, 4339-4343
6. Sono, M., Andersson, L. A., and Dawson, J. H. (1982) *J. Biol. Chem.* **257**, 8308-8320
7. Dawson, J. H., and Sono, M. (1987) *Chem. Rev.* **87**, 1255-1276

CHAPTER 3

Ferric, not ferrous, heme activates DGCR8 for primary microRNA processing

(*Proc. Natl. Acad. Sci. USA*, 109, 1919-1924.)

Ferric, not ferrous, heme activates RNA-binding protein DGCR8 for primary microRNA processing

Ian Barr[a], Aaron T. Smith[b], Yanqiu Chen[a], Rachel Senturia[a], Judith N. Burstyn[b], and Feng Guo[a,1]

[a]Department of Biological Chemistry, David Geffen School of Medicine, Molecular Biology Institute, University of California, Los Angeles, CA 90095; and [b]Department of Chemistry, University of Wisconsin, Madison, WI 53706

Edited by Harry B. Gray, California Institute of Technology, Pasadena, CA, and approved December 15, 2011 (received for review September 6, 2011)

The RNA-binding protein DiGeorge Critical Region 8 (DGCR8) and its partner nuclease Drosha are essential for processing of microRNA (miRNA) primary transcripts (pri-miRNAs) in animals. Previous work showed that DGCR8 forms a highly stable and active complex with ferric [Fe(III)] heme using two endogenous cysteines as axial ligands. Here we report that reduction of the heme iron to the ferrous [Fe(II)] state in DGCR8 abolishes the pri-miRNA processing activity. The reduction causes a dramatic increase in the rate of heme dissociation from DGCR8, rendering the complex labile. Electronic absorption, magnetic circular dichroism, and resonance Raman spectroscopies indicate that reduction of the heme iron is accompanied by loss of the cysteines as axial ligands. ApoDGCR8 dimers, generated through reduction and removal of the heme, show low levels of activity in pri-miRNA processing in vitro. Importantly, ferric, but not ferrous, heme restores the activity of apoDGCR8 to the level of the native ferric complex. This study demonstrates binding specificity of DGCR8 for ferric heme, provides direct biochemical evidence for ferric heme serving as an activator for miRNA maturation, and suggests that an intracellular environment increasing the availability of ferric heme may enhance the efficiency of pri-miRNA processing.

DiGeorge syndrome | Microprocessor | Pasha | redox | ligand switching

MicroRNAs (miRNAs) are a class of non-protein-coding RNAs about 22 nt in length (1, 2). They are involved in nearly every aspect of development and cell physiology and contribute to diseases such as cancer and DiGeorge syndrome (3–6). miRNAs are produced from long primary transcripts (pri-miRNA) that may be introns in messenger RNAs, or independent noncoding transcripts. The miRNA maturation pathway includes sequential cleavages in the nucleus and cytoplasm (7–9). DGCR8 is a RNA-binding protein that is essential for maturation of all canonical miRNAs (10–15). DGCR8 and its partner, the RNase III enzyme Drosha (16), specifically recognize and cleave pri-miRNAs in the nucleus to produce an intermediate called precursor miRNAs (pre-miRNAs). DGCR8 and Drosha copurify with each other from cell extracts and are collectively called the Microprocessor complex (10–12). Unlike Dicer, another RNase III enzyme that cleaves pre-miRNAs in the cytoplasm to generate miRNA duplexes, Drosha does not cleave pri-miRNA substrates in the absence of its RNA-binding partner DGCR8. DGCR8 makes a major contribution to the recognition of pri-miRNAs through highly cooperative binding and formation of higher-order structures (17–20).

Based on a yellow color that was associated with a recombinant human DGCR8 construct called NC1 (Fig. 1*A*), we found that DGCR8 binds heme (18). Each dimeric NC1 binds one heme molecule and heme-bound NC1 dimers are much more active in pri-miRNA processing than the heme-free monomers. This observation led us to suggest that heme is involved in regulation of miRNA maturation. To understand the potential function of heme in miRNA biogenesis, we have characterized the DGCR8–heme interaction using biochemical and structural methods. DGCR8 binds heme using a heme-binding domain (HBD) located in the central region of the 773-residue polypeptide chain, N-terminal to two double-stranded RNA-binding domains (dsRBDs) (Fig. 1*A*). The HBD of DGCR8 contains an N-terminal dimerization subdomain that is primarily composed of a WW motif. Our crystal structure of the dimerization domain demonstrates that the WW motif and its C-terminal neighboring region form an extensive dimerization interface even in the absence of heme, and this domain seems to directly contribute a surface for heme binding (21). Recently, we found that the heme bound to native NC1 is ferric, that two cysteine (Cys) side chains bind to the heme iron in a ligation configuration that has not been observed in any other heme protein, and that the NC1–ferric heme complex is highly stable (22).

Most canonical heme proteins stably bind both the ferrous and ferric forms of heme. Here we characterize the DGCR8-ferrous heme complexes, and we find that reduction of the ferric heme iron in DGCR8 greatly increases the rate of heme dissociation. Spectroscopic data show that the dual cysteine ligands of the Fe(III) heme–DGCR8 complex are lost upon reduction of the heme iron. Taking advantage of the fast dissociation of Fe(II) heme from DGCR8, we generate the apoNC1 via reduction of the heme iron and find that ferric, but not ferrous, heme activates miRNA processing activity of DGCR8 in vitro. The biological implications of these findings are discussed.

Results

Fe(II) DGCR8 is Inactive in Pri-miRNA Processing. To test the importance of the heme iron redox state on the activity of DGCR8, we reduced the native Fe(III) heme-bound NC1 dimer and examined its activity in reconstituted pri-miRNA processing assays (18). Incubation of Fe(III) NC1 with excess dithionite resulted in slow reduction of the heme iron to Fe(II), which approached completion at room temperature in 30–60 min (Fig. 1 *B* and *C*). At pH 8.0, the Fe(II) NC1 complex displayed a Soret peak at 425 nm and β-, α-bands at 530 and 557 nm, respectively; the Soret peak at 450 nm and the broad α/β-bands at 556 nm of the Fe(III) NC1 disappeared (Fig. 1*B*). At pH 8.0, the Fe(II) heme–NC1 complex tended to precipitate at concentrations higher than 2 μM. We performed the reduction at pH 6.0 and found that Fe(II) NC1 remained soluble and had an absorption maximum at 390 nm, which was partially obscured by the dithionite absorption (Fig. 1*C*). The lack of distinct features in the electronic absorption spectrum of Fe(II) NC1 at pH 6.0 raised the question as to whether the Fe(II) heme is free in solution. We disfavor this possibility because the spectrum of Fe(II) NC1 is distinct from those displayed by Fe(II) heme alone or in the presence of DGCR8$^{276–353}$, a dimerization domain-only variant that cannot

---

Author contributions: I.B., A.T.S., J.N.B., and F.G. designed research; I.B., A.T.S., Y.C., and R.S. performed research; I.B., A.T.S., J.N.B., and F.G. analyzed data; and I.B., A.T.S., J.N.B., and F.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: fguo@mbi.ucla.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1114514109/-/DCSupplemental.

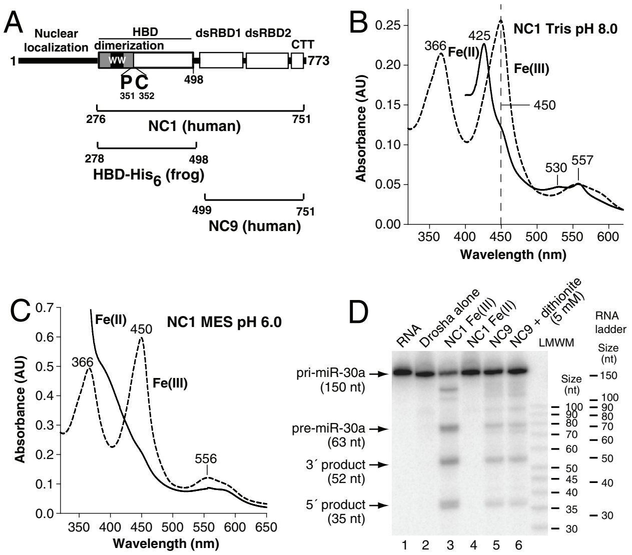

Fig. 1. Reduction of the Fe(III) heme in human NC1 diminishes pri-miRNA processing. (*A*) Domain structure of human DGCR8. The dsRBDs and C-terminal tail (CTT) are required for cooperative association with pri-miRNAs and for triggering cleavage by Drosha. The human NC1, NC9, and frog DGCR8 HBD-His$_6$ proteins used in this study are represented by brackets. (*B*) Electronic absorption spectra of NC1 (7.1 μM) recorded before (dashed line) and after (solid line) incubation with solid dithionite at 25 °C under Ar(g) for 60 min. (*C*) Electronic absorption spectra of NC1 (10 μM) reduced at pH 6.0 (50 mM MES) using 2 mM dithionite at room temperature under anaerobic conditions for 70 min. (*D*) Reconstituted pri-miR-30a processing assays were performed at 37 °C for 45 min using recombinant His$_6$-Drosha$^{390–1374}$ and various forms of DGCR8 as indicated. The Fe(III) and Fe(II) NC1 dimers were present at 25 nM and the NC9 monomer concentration was 100 nM. The relationship between the low molecular weight marker (LMWM) and the Decade RNA ladder was inferred from a comparison on a similar gel (Fig. S3).

bind heme (Fig. S1). Furthermore, the Fe(II) NC1 species at pH 6.0 and 8.0 are interconvertible (Fig. S2).

The Fe(II) NC1 complex was tested in reconstituted pri-miRNA processing assays with recombinant human Drosha$^{390–1374}$. The pri-miRNA processing reactions were performed anaerobically to avoid reoxidation of Fe(II) heme. Under these conditions, Fe(III) NC1 is highly active, whereas Fe(II) NC1 is inactive (Fig. 1*D*, lanes 3 and 4). Dithionite (1 mM) was present in the pri-miRNA processing reaction with Fe(II) NC1. To rule out the possibility that dithionite inactivated pri-miRNA processing, we added dithionite to pri-miRNA processing assays reconstituted using NC9, a DGCR8 construct that does not contain heme due to the lack of the HBD but is active in pri-miR-30a (pri-miRNA of miR-30a) processing (Fig. 1*A*). The results showed clearly that dithionite does not interfere with the pri-miRNA processing activity of Drosha and NC9 (Fig. 1*D*, lanes 5 and 6). Thus, these results indicate that, unlike the Fe(III) heme–DGCR8 complex, the Fe(II) DGCR8 complex is not active in pri-miRNA processing.

Reduction of the DGCR8–Heme Complex Greatly Increases the Rate of Heme Dissociation. To understand the mechanism of the heme-reduction-mediated inactivation, we analyzed the Fe(II) NC1 using size exclusion chromatography (SEC). The Fe(II) NC1 protein eluted at the same volume as that of the Fe(III) NC1 (Fig. 2*A*), indicating that the dimeric state was unchanged upon reduction of the heme iron. However, little absorption at 390 nm was observed in the elution peak of Fe(II) NC1, suggesting that most Fe(II) heme had dissociated from DGCR8. The possibility of reoxidized heme remaining bound to NC1 was ruled out by the lack of absorption at 450 nm in the chromatogram.

To further corroborate the instability of the Fe(II) heme–DGCR8 complex, we studied the reduction of frog DGCR8 HBD-His$_6$ (frog HBD) (Fig. 2*B*). We previously showed that the frog HBD binds Fe(III) heme in the absence of the dsRBDs and displays spectroscopic properties very similar to those of the human NC1 (22). Upon 1-h incubation with 2 mM dithionite at pH 8.0 and 37 °C, the frog HBD remained soluble, and the 366- and 450-nm split Soret peaks were replaced by a single sharp peak at 424 nm (Fig. 2*B*). The broad absorption envelope at approximately 557 nm, characteristic of the α/β region of an Fe(III) heme, was replaced by two new, distinct peaks at 529 and 558 nm, characteristic of β- and α-bands of an Fe(II) heme (Fig. 2*B*). The similar absorption peaks of Fe(II) frog HBD and Fe(II) NC1 at pH 8.0 (Figs. 1*B* and 2*B*) suggest that the environments of the heme are alike in these proteins. At pH 6.0, Fe(II) frog HBD precipitated.

We previously showed that Fe(III) NC1 was highly stable, with no heme transfer to apomyoglobin over 4 d (22). Similarly, the Fe (III) frog HBD complex did not transfer heme when incubated with apomyoglobin for 5 d (Fig. 2*C*). Apomyoglobin has an extremely high affinity for heme ($K_\mathrm{d} = 3 \times 10^{-15}$ M), making it an effective heme scavenger (23). In sharp contrast, when the Fe(II) frog HBD was incubated with a molar excess of apomyoglobin at room temperature, nearly complete transfer of Fe(II) heme from the frog HBD to myoglobin was observed within the experimental time of 1–2 min, as indicated by the shift of Soret peak from 424 to 431 nm (Fig. 2*B*). The transfer of heme from frog HBD to myoglobin was so rapid that we could not measure the rate using a stopped-flow apparatus. Therefore, we conclude that reduction

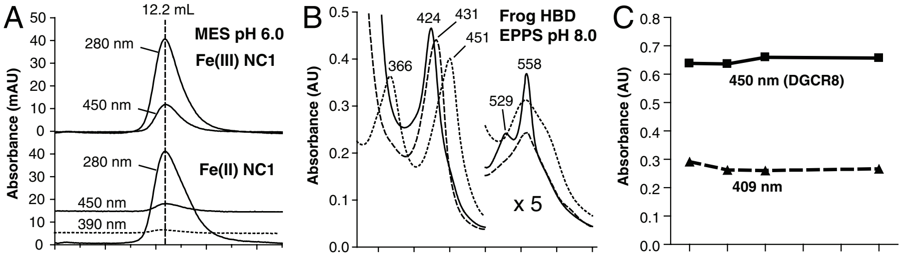

Fig. 2. Reduction of the heme iron in DGCR8 greatly decreases the stability of the heme–protein complex. (*A*) Size exclusion chromatograms of Fe(III) and Fe(II) NC1 at pH 6.0 (50 mM MES). Fe(II) NC1 was prepared as described in Fig. 1*C*. (*B*) Electronic absorption spectra of frog HBD (8 μM) at pH 8.0 [50 mM 4-(2-hydroxyethyl)-1-piperazinepropanesulfonic acid (EPPS)]. The Fe(III) form of the protein (dotted line) was reduced to the Fe(II) form (solid line) through addition of solid dithionite for 60 min. Subsequently, 66 μL of 80 μM apomyoglobin was added to 200 μL of the Fe(II) HBD solution (dashed line, normalized to compensate for dilution). (*C*) Fe(III) frog HBD (7 μM) was incubated with a sixfold excess of apomyoglobin at room temperature. The absorbance at 450 nm [Fe(III) DGCR8] and 409 nm (metmyoglobin) are plotted.

of the heme iron causes a dramatic increase of the rate of heme dissociation from DGCR8.

Reduction of the DGCR8–Heme Complex is Accompanied by Ligand Switching. Reduction of the DGCR8 heme results in the loss of the two cysteine ligands that were bound to the Fe(III) heme. When dithionite is added to Fe(III) frog HBD at pH 8.0 and 37 °C, the 451-nm peak slowly loses intensity while a sharper Soret peak at 424 nm grows in. The Soret band shift from 450 to 424 nm suggests that no cysteine(thiolate) ligand is bound to the Fe(II) DGCR8 heme, because six-coordinate low-spin thiolate-ligated Fe(II) hemes exhibit red-shifted Soret peaks (440–460 nm) (24), whereas five-coordinate high-spin thiolate-ligated Fe(II) hemes exhibit Soret peaks in the 406- to 412-nm region (25). The Fe(II) frog HBD 424-nm band has a relatively low molar absorptivity ($\varepsilon_{424} \sim 55$ mM^{-1} cm^{-1}) and the α/β region of the spectrum, although consistent with that of a low-spin Fe(II) heme, is broad. These features suggest that multiple spin and/or coordination states may be present (26). Further evidence for the formation of intermediates and/or multiple products is provided by the fact that the reduction process is not isosbestic. Fe(II) NC1 exhibits a different Soret maximum at pH 6.0 from those of Fe(II) NC1 and Fe(II) HBD at pH 8.0, indicating that the heme spin and coordination state changes with pH.

Magnetic circular dichroism (MCD) spectroscopy further supports the conclusion that the Fe(II) frog HBD has lost the bis-Cys ligation present in the Fe(III) complex and exists as a mixture of spin and coordination states at pH 8.0. The α/β region of the Fe(II) frog HBD MCD spectrum is dominated by a derivative-shaped, temperature-independent *A* term with a cross-over position of 558 nm (α-band), consistent with the presence of a six-coordinate, low-spin Fe(II) heme (Fig. 3*A*). However, the Soret region of the Fe(II) frog HBD MCD spectrum is dominated by an inverted *C* term with trough-cross-over-peak positions of 422–431–442 nm, consistent with the presence of a five-coordinate, high-spin Fe(II) heme. The magnetic saturation behavior of the most intense peak of the *C* term, 442 nm (Fig. 3*A*, *Inset*), taken at 2.5, 4.0, 8.0, 15, and 25 K has a characteristic shape and nonoverlapping nature that further confirm the presence of high-spin, $S = 2$, Fe(II) heme. At 50 K, the MCD spectrum of Fe(II) frog HBD shows greater intensity in the Soret *C* term than the *A* term in the α/β region, implying that high-spin Fe(II) heme is the major species at pH 8.0 (27). The electronic absorption and MCD peak positions and intensities of Fe(II) DGCR8 do not match those of thiolate-bound Fe(II) human cystathionine β-synthase or thiol-bound Fe(II) myoglobin H93G, suggesting that Cys is not bound to the Fe(II) heme of DGCR8 (28, 29).

Resonance Raman spectroscopy reinforces the conclusion that the heme in Fe(II) frog HBD exists as a mixture of spin and coordination states. The most intense oxidation state marker band, ν_4, displays a dramatic downshift in energy from 1,372 to 1,358 cm^{-1}, consistent with the reduction of Fe(III) to Fe(II) heme (30, 31) (Fig. 3*C*). The most prominent spin and coordination state marker band, ν_3, shifts from 1,501 cm^{-1} upon heme reduction and becomes split between 1,470 cm^{-1} (high-spin, five-coordinate) and 1,489 cm^{-1} (low-spin, six-coordinate), implicating a mixture of spin and coordination states for the Fe(II) heme in frog HBD (31, 32). At higher energy, the oxidation and coordination state marker bands ν_2 and ν_{10} are identified at 1,579 and 1,615 cm^{-1}, respectively, similar to those of other Fe(II) heme proteins that have mixed coordination and ligation states (30–32). Thus, from the electronic absorption, MCD, and resonance Raman data we conclude that, when the DGCR8 heme is reduced, the two cysteine ligands are lost, and one or two new ligands bind to the heme. The exact nature of the interaction between Fe(II) heme and DGCR8 (i.e. the identity of the new ligands and the specificity of the Fe(II) heme–protein interaction) will be reported elsewhere.

Association of ApoNC1-P351A with Fe(II) Heme. Neither the wild-type human NC1 nor the frog HBD can be expressed in dimeric heme-free (apo) forms. However, the NC1-P351A mutant has reduced affinity for heme and can be expressed in *Escherichia coli* as a heme-free dimer. The purified apoNC1-P351A dimer is very soluble at pH 8.0 and binds Fe(III) heme in vitro to reconstitute the complex with its characteristic absorption peaks at 365, 447, and 556 nm (22). Here we incubated heme-free NC1-P351A dimer with various molar equivalents of Fe(II) heme at pH 8.0 under anaerobic conditions; the electronic absorption spectra displayed peaks at 424, 529, and 557 nm, respectively, which are very close to those of reduced NC1 (Fig. 1*B* and Fig. S4*A*). This result suggests that an Fe(II) heme–NC1-P351A complex has been reconstituted. SEC analysis showed the reconstituted Fe(II)

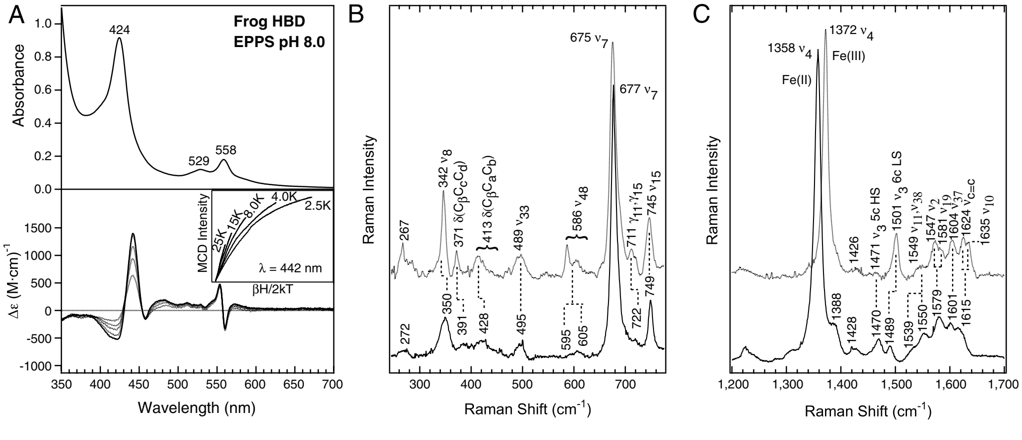

Fig. 3. The Fe(II) heme in frog HBD is a mixture of five-coordinate, high-spin and six-coordinate, low-spin species without a cysteine thiolate ligand. (*A*) The electronic absorption spectrum (*Upper*) of Fe(II) frog HBD (11.7 μM) in 50 mM 4-(2-hydroxyethyl)-1-piperazinepropanesulfonic acid (EPPS) (pH 8.0), 400 mM NaCl, and 1 mM sodium dithionite at 37 °C. MCD spectra (*Lower*) of the Fe(II) frog HBD (29.4 μM) in 20 mM EPPS (pH 8.0), 160 mM NaCl, 60% (vol/vol) glycerol, and 7 mM sodium dithionite at 4, 8, 15, 25, and 50 K. (*Inset*) The magnetic saturation behavior of the MCD *C*-term intensity at 442 nm taken at 2.5, 4.0, 8.0, 15, and 25 K. (*B* and *C*) Resonance Raman spectra are shown for Fe(III) (dotted line) and Fe(II) (solid line) frog HBD for low energy (*B*) and high energy (*C*) regions. Fe (III) HBD (153 μM) and Fe(II) HBD (136 μM) samples were in 45 mM EPPS (pH 8.0), 360 mM NaCl, and 10% (vol/vol) glycerol; Fe(II) HBD also contained 7 mM sodium dithionite. For the Fe(III) protein, spectra were acquired by excitation with a 457.9-nm line with 20 mW of power at the sample; for the Fe(II) protein, spectra were acquired by excitation with a 413.1-nm line with 14 mW of power at the sample.

NC1-P351A complex is labile (Fig. S4*B*), similar to the Fe(II) wild-type NC1.

ApoNC1 Produced Through Reduction Binds Both Fe(III) and Fe(II) Heme in Vitro. To directly test the effect of heme binding on the activity of DGCR8 in vitro, it is desirable to generate heme-free wild-type protein. To accomplish this, we took advantage of the labile nature of the Fe(II) NC1 complex. The heme-bound NC1 dimer was reduced at pH 6.0 using dithionite, and the Fe(II) heme was removed through incubation with apomyoglobin followed by SEC (Fig. 4*A*). The A_{280}/A_{450} ratio of the NC1 protein increased from 2.8 to 27 through this procedure, indicating that approximately 90% of heme was removed. The small amount of heme that remained bound to NC1 contained Fe(III). Incubation of the apoNC1 protein with equimolar Fe(III) heme restored the 366-, 450-, and 556-nm peaks (Fig. 4*A*), indicating that the Fe(III) heme–NC1 complex was successfully reconstituted. The 366- and 450-nm peaks of the reconstituted Fe(III) NC1 complex are of nearly equal intensities. In contrast, the 366-nm peak of the native NC1 complex is always less intense than the 450-nm peak. This observation indicates that the reconstituted complex may have subtle differences relative to the native complex. Incubation of Fe(II) heme with apoNC1 at pH 6.0 produced a complex with an absorption peak at approximately 390 nm (Fig. 4*A*), similar to that of the reduced NC1 (Fig. 1*C*).

Fe(III), not Fe(II), Heme Activates DGCR8 for Pri-miRNA Processing. We tested the activity of the reconstituted Fe(III) and Fe(II) heme NC1 complexes using in vitro pri-miRNA processing assays (Fig. 4 *B–D*). Three pri-miRNA fragments, pri-miR-30a, pri-miR-21, and pri-miR-380, were used in these assays. The native NC1 complex was much more active than apoNC1 (lanes 3 and 4, 10 and 11, 17 and 18). The reconstituted Fe(III) NC1 complex was as active as native NC1 (lanes 5, 12, and 19), whereas the reconstituted Fe(II) NC1 complex was as inactive as the apoNC1 (lanes 6, 13, and 20). To prevent oxidation of Fe(II) heme, these assays were performed under anaerobic (N_2 gas) conditions. As controls, we also performed pri-miRNA processing assays using native NC1 under aerobic conditions (lanes 7 and 14) and observed activity indistinguishable from that in anaerobic conditions (lanes 3 and 10). This result suggests that the oxygen in the air does not affect the activity of Fe(III) NC1. Overall, these experiments clearly demonstrate that heme-free DGCR8 has very low pri-miRNA processing activity, and that Fe(III) heme activates DGCR8, whereas Fe(II) heme does not.

ApoNC1 Binds Pri-miRNAs with Reduced Cooperativity. To understand why the apoNC1 dimer is defective in pri-miRNA processing, we used a filter-binding assay to examine its interaction with pri-miR-30a. The binding data were best fit using a cooperative dimer model, in which two apoNC1 dimers bind cooperatively to a pri-miR-30a RNA, with a $\sqrt{K_d}$ of 6.0 ± 1.6 nM (Fig. 4*E*) and a Hill coefficient (n) of 1.6 ± 0.2 (Fig. 4*F*). Our control experiments and previous work showed that Fe(III) heme-bound NC1 binds pri-miR-30a with similar affinities, but with a higher Hill coefficient (*ca.* 3) (18). Thus, heme removal from DGCR8 likely inhibits pri-miRNA processing via reducing the binding cooperativity and hampering the formation of productive higher-order structures. This observation is consistent with our recent report demonstrating that DGCR8 recognizes pri-miRNAs through highly cooperative binding and formation of higher-order (trimer of dimers) structures (17).

Discussion

This study advances our understanding of the function of heme in miRNA maturation by demonstrating that Fe(III) heme binds to the apoNC1 dimer to activate its pri-miRNA processing activity, and that DGCR8 has an unusual specificity for Fe(III) heme over Fe(II) heme. In our initial discovery of the DGCR8–heme interaction, we observed that heme-bound NC1 dimer is much more

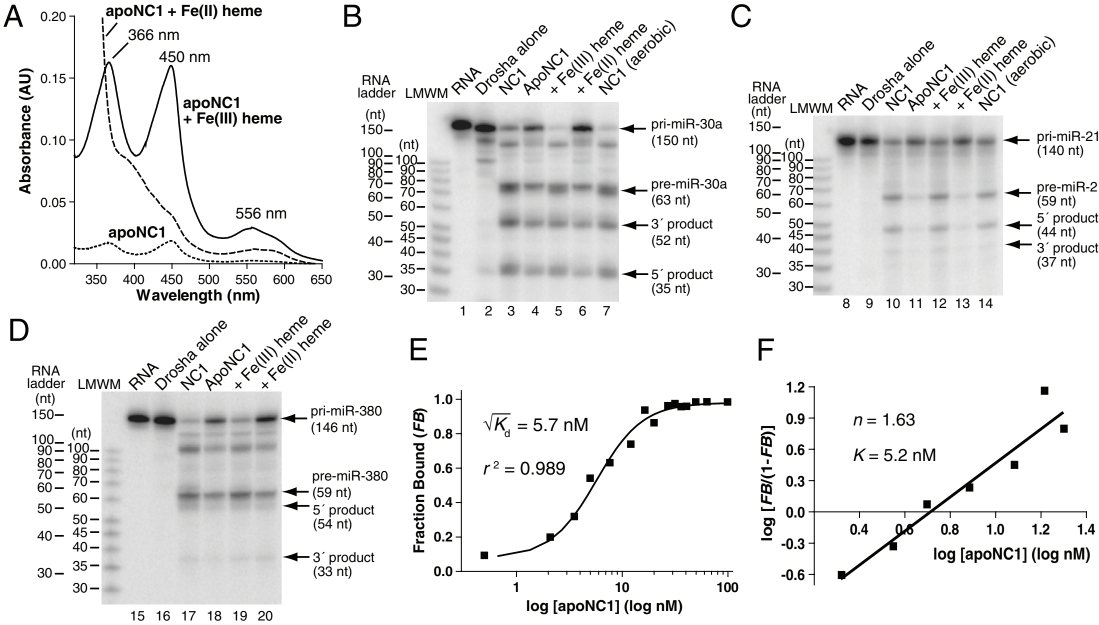

Fig. 4. Fe(III), not Fe(II), heme reconstitutes DGCR8 for pri-miRNA processing in vitro. (*A*) Electronic absorption spectra of apoNC1 (dotted line), prepared as described in *Materials and Methods*, and after incubation with equimolar Fe(III) (solid line) or Fe(II) (dashed line) heme. The protein solutions contained 50 mM MES (pH 6.0), 400 mM NaCl, and 1 mM DTT. (*B–D*) Uniformly ^{32}P-labeled pri-miR-30a (*B*), pri-miR-21 (*C*), and pri-miR-380 (*D*) were incubated at 37 °C for 45 min, with either Drosha$^{390-1374}$ alone (lanes 2, 9, and 16) or Drosha$^{390-1374}$ together with various forms of NC1 as indicated (lanes 3–7, 10–14, and 17–20). All reactions were performed in anaerobic conditions, except the ones analyzed in lanes 7 and 14. (*E*) Filter-binding assays demonstrate that apoNC1 associates with pri-miR-30a with lower cooperativity than that of the Fe(III) heme-bound NC1 (18). The data were best fit using a cooperative dimer model. (*F*) Hill plot of data from *E* around the binding transition. The result shown is one of three experiments. K is defined as $10^{(x\text{-intercept})}$. Low molecular weight marker, LMWM.

active than the heme-free monomer in reconstituted pri-miRNA processing reactions (18). However, we could neither add heme to the heme-free monomer to activate NC1 nor remove heme from the heme-bound dimer to inactivate it. Later, we realized that the existence of the NC1 "monomer" is at least partially attributable to a NC1 heterodimer in which one subunit is cleaved during bacterial overexpression and purification (22). Characterization of the Fe(II) DGCR8 complexes led to the successful removal of heme from NC1 via reduction to produce the apoNC1 dimer, which allowed us to directly link heme binding with pri-miRNA processing. Importantly, using the apo-NC1 dimer, we were able to show definitively that only Fe(III) heme, and not Fe(II) heme, activates NC1 for pri-miRNA processing.

DGCR8 is a heme protein with high specificity for binding Fe(III) heme. Previous studies showed that the association rates (k_{on}) for heme binding to globin variants, BSA, and several other heme proteins are similar to each other ($1\text{–}10 \times 10^7\ \mathrm{M}^{-1}\,\mathrm{s}^{-1}$) and are independent of protein structure (33, 34). Furthermore, the k_{on} values of Fe(III) heme for apomyoglobin and BSA, as measured using fluorescence quenching, are almost the same as those of Fe(II) heme (33). Thus, the affinities of these proteins for heme, which vary up to 10^6-fold, are primarily determined by the rates of heme dissociation (k_{off}). The estimated k_{off} for the DGCR8–heme complex increased from $<2 \times 10^{-6}\ \mathrm{s}^{-1}$ for the Fe(III) complex ($t_{1/2} \gg 5$ d assuming that the K_{d} is higher than that of myoglobin, $k_{\mathrm{off}} = \ln 2/t_{1/2}$) to $>1 \times 10^2\ \mathrm{s}^{-1}$ (too fast to be measured using the stopped-flow method). The $>10^7$-fold difference in k_{off} of DGCR8 for Fe(III) and Fe(II) heme suggests that the thermodynamic stability of Fe(III) complex is much greater than that of the Fe(II) complex.

The specificity of DGCR8 for Fe(III) heme is likely contributed by the use of two cysteine side chains as axial ligands. The affinity of a typical heme protein for heme is contributed by the axial ligands that bind to the iron, a hydrophobic pocket that surrounds the porphyrin ring, and amino acid residues that make specific interactions with the heme periphery, including salt bridges to the heme propionates. In myoglobin, mutation of the proximal histidine ligand to glycine reduces the affinity for heme by a factor of 10^4, which provides an estimate of the contribution for heme binding by this axial ligand (33). Mutation of cysteine-352 in both NC1 subunits completely abolishes heme binding during bacterial expression, demonstrating the importance of dual cysteine axial ligation in the binding of Fe(III) heme to DGCR8 (18). Further evidence for the importance of dual cysteine ligation is provided by the $>10^7$-fold increase of k_{off} for heme when the DGCR8 heme is reduced and both cysteine ligands are lost. That Fe(II) heme remains bound to DGCR8, albeit more weakly, is presumably attributable to other DGCR8–heme interactions and to the new ligand or ligands that replace the cysteines.

Reducing the heme abrogates function, presumably due to structural changes in the DGCR8 protein arising from ligand switching at the heme. Previously reported data suggest that at least one of the two Fe(III) heme-bound cysteine ligands, and possibly both, is deprotonated (22). When the heme is reduced from Fe(III) to Fe(II), unfavorable electrostatic interaction between the more electron rich Fe(II) and the anionic thiolate ligands drives them both off the heme. Interestingly, we see no evidence for DGCR8-bound heme with only a single cysteine ligand, either in the Fe(III) or Fe(II) forms. During the slow reduction of Fe(III) to Fe(II) frog HBD, no five-coordinate thiolate-ligated intermediate was observed. Similarly, when Fe(III) human DGCR8 HBD was titrated with up to 2 molar equivalents of methylmercury, a linear decrease of the absorption peak at 450 nm was observed with no intermediate observed that could be attributable to single-cysteine-ligated species (22).

The activity of DGCR8 is dependent on the redox state of its heme cofactor in vitro, suggesting that heme may be a redox-sensitive regulator of cellular RNA processing. One working model regarding the biological function of the heme–DGCR8 interaction is that heme may serve as a ligand of DGCR8 to activate pri-miRNA processing (the heme sensing model). The weak pri-miRNA processing activity of the heme-free NC1 dimer and the activation by Fe(III) heme strongly support this model. The lack of activation of DGCR8 by Fe(II) heme further suggests that pri-miRNA processing might be regulated specifically by the availability of Fe(III) heme (the ferric heme sensing model). Heme is produced in the Fe(II) form, and the last step of the heme biosynthesis pathway, incorporation of Fe(II) into protoporphyrin IX by ferrochelatase, occurs on the mitochondrial inner membrane in animal cells (35). Little is known about how heme is transported from the site of Fe(II) insertion to where it is incorporated into host proteins, or how the oxidation state of heme may be altered during trafficking (36). Our data suggest that an intracellular environment favoring conversion of Fe(II) heme to Fe(III) heme may increase the efficiency of pri-miRNA processing.

The results reported herein suggest an opportunity for controlling pri-miRNA processing in therapeutics. DGCR8 is among the 35–60 genes at the 22q11.2 locus that are heterozygously deleted in DiGeorge syndrome (37, 38). Haploinsufficiency of DGCR8 in mouse models results in reduced processing of a subset of miRNAs in the brain and in deficits in synaptic plasticity in the prefrontal cortex that are associated with symptoms of the syndrome (4–6). The unique heme-binding properties revealed in our studies suggest that it may be possible to alter pri-miRNA processing efficiency using heme derivatives without affecting other heme proteins. This approach may be used for correcting miRNA processing defects in DiGeorge syndrome and other diseases without genetic manipulation.

Materials and Methods

Expression, Purification, Reduction, Heme Removal, and Reconstitution of DGCR8 Proteins. The human NC1 (heme-bound wild-type and heme-free P351A mutant), DGCR8$^{276\text{–}353}$, and frog HBD-His$_6$ proteins were expressed in *E. coli* and purified as previously described (18, 21, 22). The purified proteins were in 20 mM Tris (pH 8.0), 400 mM NaCl, and 1 mM DTT, and in some spectroscopic studies were exchanged into desired pH buffers, as indicated in the figures and figure legends, using centrifugal concentrators. Reduction of heme-bound NC1 and frog HBD was performed in an anaerobic chamber filled with N_2(g), unless stated otherwise. To produce apoNC1, freshly purified Fe(III) heme-bound NC1 was reduced at pH 6.0 with 2 mM sodium dithionite at room temperature in an anaerobically sealed cuvette until the 450-nm absorbance peak disappeared (*ca.* 1 h). Excess apomyoglobin was added to scavenge Fe(II) heme dissociated from NC1, and the proteins were separated using SEC at pH 6.0, as described below. For reconstitution experiments, hemin chloride was dissolved in 1.4 M NaOH at 100 mM, incubated at room temperature for approximately 30 min, and then diluted in water to give a 100-μM stock solution. An Fe(II) heme stock solution (100 μM) was prepared by addition of 5 mM dithionite to a hemin solution. Sodium dithionite (1 mM) was included in the apoNC1 and apoNC1-P351A protein solutions during reconstitution with Fe(II) heme.

Electronic Absorption Spectroscopy. Electronic absorption spectra were recorded using either a DU800 spectrophotometer (Beckman-Coulter; bandwidth ≤1.8 nm) (Figs. 1, 2, and 4, and Figs. S1, S2, and S4*A*) or a double-beam Varian Cary 4 Bio spectrophotometer with a temperature controller (Agilent Technologies) with its spectral bandwidth set to 0.5 nm (Fig. 3). The samples containing Fe(II) heme were analyzed in sealed cuvettes under N_2 or Ar gas.

Size Exclusion Chromatography. The SEC analysis was performed at room temperature using a Superdex-200 10/300 GL column (GE Healthcare). The running buffer contained 400 mM NaCl, 1 mM DTT, and either 50 mM MES (pH 6.0) (Fig. 2*A* and in production of apoNC1) or 20 mM Tris (pH 8.0) (Fig. S4*B*), and was always degassed.

Expression and Purification of His$_6$-Drosha$^{390\text{–}1374}$. The His$_6$-Drosha$^{390\text{–}1374}$ protein was overexpressed in Sf9 insect cells using the baculovirus system and was purified using Ni affinity chromatography. Details may be found in *SI Text*.

In Vitro Pri-miRNA Binding and Processing Assays. The assays were carried out at pH 8.0 as described previously (18), with adaptation for anaerobic conditions where indicated. Details may be found in *SI Text*.

MCD Spectroscopy. MCD spectra were recorded and analyzed as described previously (22). The Fe(II) frog HBD (*ca.* 29 μM) was prepared from the Fe(III) protein at 37 °C via addition of 7 mM sodium dithionite. Samples were transferred via gas-tight syringes into cells purged with Ar(g), flash-frozen, and stored in $N_2(l)$.

Resonance Raman Spectroscopy. Resonance Raman spectra were obtained with an excitation wavelength of either 413.1 nm from a Coherent I-302C Kr^+ laser or 457.9 nm from a Coherent I-305 Ar^+ laser in a backscattering 135° sample geometry. An Acton Research triple monochromator was used with a grating of 2,400 grooves/mm. Low incident laser powers of $\leq$20 mW were focused with a cylindrical lens onto the sample. A Princeton Instruments Spex 1877 triple spectrograph outfitted with a cooled, intensified diode array detector was operated under computer control. Samples were placed in a quartz Dewar filled with ice water to reduce local heating. Peak positions were calibrated relative to a Na_2SO_4 standard peak at 983 cm^{-1}.

ACKNOWLEDGMENTS. We thank J. Valentine and K. Barnese for use of the anaerobic chamber, J. Feigon and T.C. Brunold for use of spectrophotometers, S. Weitz for sharing reagents, and M. Phillips in the University of California, Los Angeles–Department of Energy Biochemistry Instrumentation Facility for technical support. This project is supported by National Institutes of Health Grants GM080563 (to F.G.) and HL065217 (to J.N.B.).

1. Ambros V (2004) The functions of animal microRNAs. *Nature* 431:350–355.
2. Bartel DP (2009) MicroRNAs: Target recognition and regulatory functions. *Cell* 136:215–233.
3. Croce CM (2009) Causes and consequences of microRNA dysregulation in cancer. *Nat Rev Genet* 10:704–714.
4. Stark KL, et al. (2008) Altered brain microRNA biogenesis contributes to phenotypic deficits in a 22q11-deletion mouse model. *Nat Genet* 40:751–760.
5. Fenelon K, et al. (2011) Deficiency of Dgcr8, a gene disrupted by the 22q11.2 microdeletion, results in altered short-term plasticity in the prefrontal cortex. *Proc Natl Acad Sci USA* 108:4447–4452.
6. Schofield CM, et al. (2011) Monoallelic deletion of the microRNA biogenesis gene Dgcr8 produces deficits in the development of excitatory synaptic transmission in the prefrontal cortex. *Neural Dev* 6:11.
7. Kim VN, Han J, Siomi MC (2009) Biogenesis of small RNAs in animals. *Nat Rev Mol Cell Biol* 10:126–139.
8. Faller M, Guo F (2008) MicroRNA biogenesis: There's more than one way to skin a cat. *Biochim Biophys Acta* 1779:663–667.
9. Sashital DG, Doudna JA (2010) Structural insights into RNA interference. *Curr Opin Struct Biol* 20:90–97.
10. Gregory RI, et al. (2004) The Microprocessor complex mediates the genesis of microRNAs. *Nature* 432:235–240.
11. Denli AM, Tops BB, Plasterk RH, Ketting RF, Hannon GJ (2004) Processing of primary microRNAs by the Microprocessor complex. *Nature* 432:231–235.
12. Han J, et al. (2004) The Drosha-DGCR8 complex in primary microRNA processing. *Genes Dev* 18:3016–3027.
13. Landthaler M, Yalcin A, Tuschl T (2004) The human DiGeorge syndrome critical region gene 8 and its D. melanogaster homolog are required for miRNA biogenesis. *Curr Biol* 14:2162–2167.
14. Wang Y, Medvid R, Melton C, Jaenisch R, Blelloch R (2007) DGCR8 is essential for microRNA biogenesis and silencing of embryonic stem cell self-renewal. *Nat Genet* 39:380–385.
15. Yi R, et al. (2009) DGCR8-dependent microRNA biogenesis is essential for skin development. *Proc Natl Acad Sci USA* 106:498–502.
16. Lee Y, et al. (2003) The nuclear RNase III Drosha initiates microRNA processing. *Nature* 425:415–419.
17. Faller M, et al. (2010) DGCR8 recognizes primary transcripts of microRNAs through highly cooperative binding and formation of higher-order structures. *RNA* 16:1570–1583.
18. Faller M, Matsunaga M, Yin S, Loo JA, Guo F (2007) Heme is involved in microRNA processing. *Nat Struct Mol Biol* 14:23–29.
19. Han J, et al. (2006) Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell* 125:887–901.
20. Sohn SY, et al. (2007) Crystal structure of human DGCR8 core. *Nat Struct Mol Biol* 14:847–853.
21. Senturia R, et al. (2010) Structure of the dimerization domain of DiGeorge Critical Region 8. *Protein Sci* 19:1354–1365.
22. Barr I, et al. (2011) DiGeorge Critical Region 8 (DGCR8) is a double-cysteine-ligated heme protein. *J Biol Chem* 286:16716–16725.
23. Hargrove MS, et al. (1994) Stability of myoglobin: A model for the folding of heme proteins. *Biochemistry* 33:11767–11775.
24. Dawson JH, Andersson LA, Sono M (1983) The diverse spectroscopic properties of ferrous cytochrome P-450-CAM ligand complexes. *J Biol Chem* 258:13637–13645.
25. Sono M, Stuehr DJ, Ikeda-Saito M, Dawson JH (1995) Identification of nitric oxide synthase as a thiolate-ligated heme protein using magnetic circular dichroism spectroscopy Comparison with cytochrome P-450-CAM and chloroperoxidase. *J Biol Chem* 270:19943–19948.
26. Adar F (1978) Electronic absorption spectra of hemes and hemoproteins. *The Porphyrins*, ed D Dolphin (Academic, New York), IIIA, pp 167–210.
27. Alberta JA, Andersson LA, Dawson JH (1989) Spectroscopic characterization of secondary amine mono-oxygenase Comparison to cytochrome P-450 and myoglobin. *J Biol Chem* 264:20467–20473.
28. Pazicni S, et al. (2004) The redox behavior of the heme in cystathionine beta-synthase is sensitive to pH. *Biochemistry* 43:14684–14695.
29. Perera R, et al. (2003) Neutral thiol as a proximal ligand to ferrous heme iron: implications for heme proteins that lose cysteine thiolate ligation on reduction. *Proc Natl Acad Sci USA* 100:3641–3646.
30. Spiro TG, Strekas TC (1974) Resonance Raman spectra of heme proteins. Effects of oxidation and spin state. *J Am Chem Soc* 96:338–345.
31. Hu S, Smith KM, Spiro TG (1996) Assignment of protoheme resonance Raman spectrum by heme labeling in myoglobin. *J Am Chem Soc* 118:12638–12646.
32. Tomita T, Gonzalez G, Chang AL, Ikeda-Saito M, Gilles-Gonzalez MA (2002) A comparative resonance Raman analysis of heme-binding PAS domains: Heme iron coordination structures of the BjFixL, AxPDEA1, EcDos, and MtDos proteins. *Biochemistry* 41:4819–4826.
33. Hargrove MS, Barrick D, Olson JS (1996) The association rate constant for heme binding to globin is independent of protein structure. *Biochemistry* 35:11293–11299.
34. Igarashi J, et al. (2008) The roles of thiolate-heme proteins, other than the P450 cytochromes, in the regulation of heme-sensor proteins. *Acta Chim Slov* 55:67–74.
35. Dailey HA (2002) Terminal steps of haem biosynthesis. *Biochem Soc Trans* 30:590–595.
36. Schultz IJ, Chen C, Paw BH, Hamza I (2010) Iron and porphyrin trafficking in heme biogenesis. *J Biol Chem* 285:26753–26759.
37. Karayiorgou M, Simon TJ, Gogos JA (2010) 22q11.2 microdeletions: Linking DNA structural variation to brain dysfunction and schizophrenia. *Nat Rev Neurosci* 11:402–416.
38. Shiohama A, Sasaki T, Noda S, Minoshima S, Shimizu N (2003) Molecular cloning and expression analysis of a novel gene DGCR8 located in the DiGeorge syndrome chromosomal region. *Biochem Biophys Res Commun* 304:184–190.

Supporting Information

Barr et al. 10.1073/pnas.1114514109

SI Materials and Methods

Materials. Sodium dithionite ($Na_2S_2O_4$) was purchased from Fluka and EMD Chemicals. The apomyoglobin was from Sigma-Aldrich.

Cloning, Expression and Purification of His$_6$-Drosha$^{390–1374}$. The sequence of human Drosha was subcloned into the pFastBac-HTb vector (Invitrogen) using PCR methods. The His$_6$-Drosha$^{390–1374}$ protein was overexpressed in Sf9 insect cells using the baculovirus system. The His$_6$-Drosha$^{390–1374}$ protein was extracted from the cells via sonication in a lysis buffer containing 20 mM Tris (pH 8.0), 500 mM NaCl, 10 mM imidazole, and 20% glycerol. This protein was purified using a Ni Sepharose high-performance column (GE Healthcare) and was eluted in the same buffer except containing 200 mM imidazole. The purified His$_6$-Drosha$^{390–1374}$ was dialyzed against a storage buffer containing 20 mM Tris (pH 8.0), 100 mM NaCl, 0.2 mM EDTA, and 10% (vol/vol) glycerol, and was stored at −80 °C until use.

In Vitro MicroRNA (miRNA) Primary Transcripts (pri-miRNAs) Processing Assays. The assays were carried out as described previously (1), with adaptation for anaerobic conditions. Reaction solutions containing recombinant His$_6$-Drosha$^{390–1374}$ and various forms of recombinant human RNA-binding protein DiGeorge Critical Region 8 (DGCR8) construct called NC1 were prepared in 0.3-mL tubes with silicone septa (National Scientific) in the anaerobic chamber. Heme was added to the appropriate protein solutions and mixed, and the samples were transferred to plastic tubes with screw-on silicone septa. These samples were then removed from the anaerobic chamber, and ^{32}P-labeled pri-miRNA was injected using a gas-tight syringe (Hamilton). The reactions were then incubated at 37 °C for 45 min, and 2× RNA loading dye containing 1× TBE (89 mM Tris, 89 mM boric acid, 2 mM EDTA, pH 8.3), 10 M urea, 10 mM EDTA, and 0.002% xylene cyanol was added to stop the reaction. The samples were analyzed using 15% polyacrylamide gels containing 7 M urea and autoradiography.

Stopped-Flow Measurement of Fe(II) Heme Dissociation from Frog Heme-Binding Domain. Data were collected on an SX.18MV stopped-flow spectrophotometer (Applied Photophysics) set on absorbance mode, with a 1-cm path length. The temperature of the sample chamber was set at 25 °C.

1. Faller M, Matsunaga M, Yin S, Loo JA, Guo F (2007) Heme is involved in microRNA processing. *Nat Struct Mol Biol* 14:23–29.

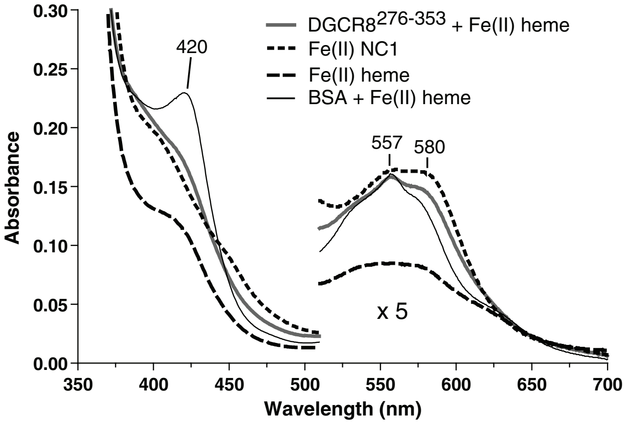

Fig. S1. The electronic absorption spectrum of Fe(II) NC1 is distinct from that of free Fe(II) heme. All solutions contained 50 mM MES (pH 6.0), 400 mM NaCl, and 5 mM sodium dithionite. Spectra were recorded of Fe(II) heme (5 μM) alone (dashed line), or in the presence of equimolar DGCR8$^{276–353}$ dimer (thick gray line) and BSA, after incubation at room temperature for 5 min. BSA is known to bind Fe(II) heme and was used as a control (thin solid line).

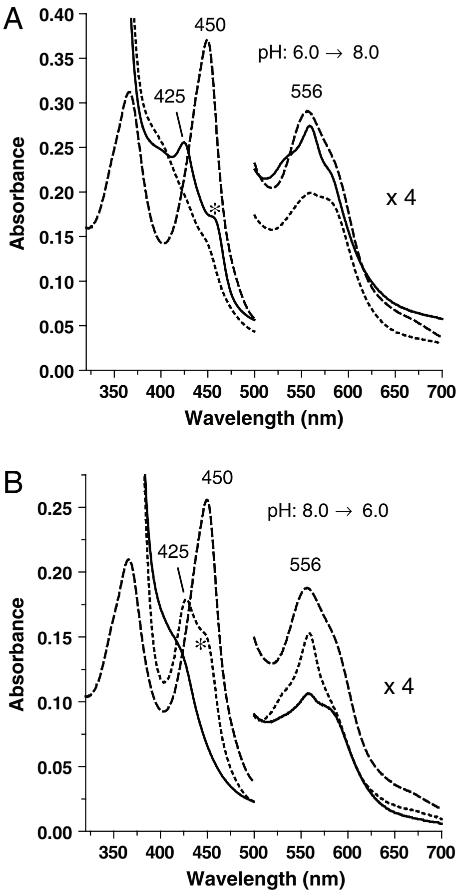

Fig. S2. The pH-dependent absorption spectral changes of Fe(II) NC1 are reversible. (*A*) Electronic absorption spectra of Fe(II) NC1 (dotted line) prepared via reduction of Fe(III) NC1 (dashed line) in 25 mM MES (pH 6.0), followed by addition of 1 M 4-(2-hydroxyethyl)-1-piperazinepropanesulfonic acid (EPPS) (pH 8.0) to a final concentration of 100 mM (solid line). The reduction was performed using 5 mM sodium dithionite under anaerobic conditions at 37 °C for 90 min. The addition of the second buffer for pH shifting was done under a stream of argon and the dilution effect on heme and protein was corrected via normalization of the absorption values in the spectra. Asterisk denotes a minor band likely due to some reoxidation of the Fe(II) heme. (*B*) Electronic absorption spectra of Fe(II) NC1 (dotted line) prepared via reduction of Fe(III) NC1 (dashed line) in 25 mM EPPS (pH 8.0), followed by addition of 100 mM MES (pH 6.0) (solid line). The experiments were performed similarly to *A*, except that reduction was achieved via incubation at room temperature for 40 min. Asterisk denotes a minor band likely due to partial reoxidation of the Fe(II) heme.

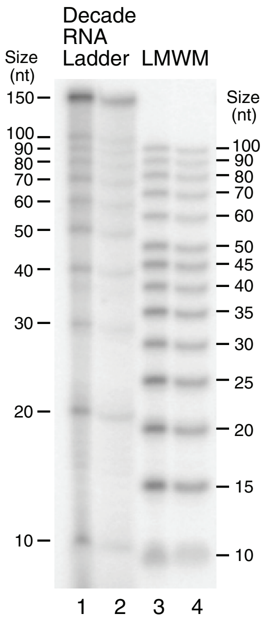

Fig. S3. Comparison of the Decade RNA ladder (Applied Biosystems) (lanes 1 and 2) and the low molecular weight marker (LMWM; Affymetrix) (lanes 3 and 4) using a 7 M urea 15% polyacrylamide gel. Ladders for lanes 1 and 3 were diluted in a formamide loading buffer, which contained 70% formamide, 0.1 M EDTA, and 1× TBE. Ladders for lanes 2 and 4 were diluted in a urea loading buffer containing 10 M urea, 20 mM EDTA, and 1× TBE. All samples were heated at 95 °C for 3 min prior to loading. The LMWM runs slower than the Decade RNA ladder, which indicates more accurately the sizes of RNAs but has become unavailable to us due to Drug Enforcement Administration regulations.

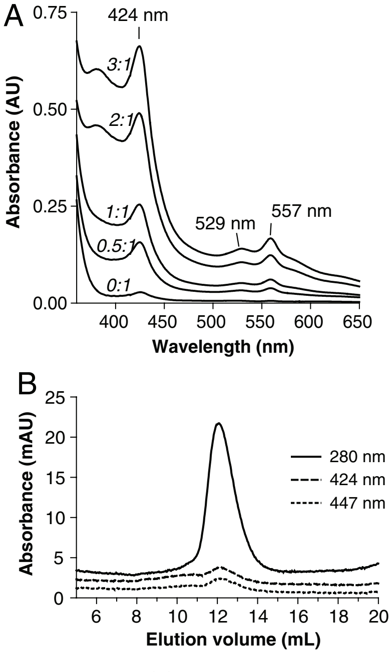

Fig. S4. Fe(II) heme forms a labile complex with apoNC1-P351A dimer. (*A*) Electronic absorption spectra of apoNC1-P351A dimer (6 μM) titrated with Fe(II) heme at pH 8.0 under anaerobic conditions. The protein solution contained 20 mM Tris (pH 8.0), 400 mM NaCl, 1 mM DTT, and 1 mM sodium dithionite. The Fe (II) heme stock solution (100 μM) was prepared through incubation of hemin solution with 5 mM sodium dithionite. The molar ratios of heme and P351A dimer are indicated in italics. As the ratio of Fe(II) heme and NC1-P351A dimer increased to over 1 : 1, a peak at 380 nm started to appear, which is likely due to nonspecifically bound heme. (*B*) Size exclusion chromatogram of the 1 : 1 Fe(II) heme:NC1-P351A complex prepared as described above. The NC1-P351A protein eluted in the absence of heme—little absorption was observed at 424 nm [the Soret peak of the Fe(II) complex] or 447 nm [the peak of the Fe(III) complex].

CHAPTER 4

Co(III)PPIX activates DGCR8 for microRNA processing

Abstract

DGCR8 is a protein component of the Microprocessor complex, which is most notably involved in one of the early steps of the microRNA biogenesis pathway. In this step, primary microRNA is cleaved to precursor microRNA by the RNAse III family member Drosha. Drosha and DGCR8 are the only Microprocessor components known to be necessary for pri-miRNA cleavage, and can be used to reconstitute the reaction *in vitro*. Previous work in our lab has shown that DGCR8 is a heme-binding protein. The heme was found to be in the Fe(III) state in the native complex, and the iron is ligated by the sulfurs of two cysteine residues, one from each subunit of the homodimer. Here we expand the repertoire of potential DGCR8 cofactors, and show interactions with two other metallo-protoporphyrin IX derivatives, Co(III)PPIX and Cr(III)PPIX, out of a total of 10 compounds screened. We show that Co(III)PPIX is able to activate primary microRNA processing, but Cr(III)PPIX is not. The potential research and therapy uses of Co(III)PPIX and similar derivatives are discussed.

Abbreviations Used

PPIX, protoporphyrin IX; miRNA, microRNA; DGCR8, DiGeorge Critical Region 8; pri-miRNA, primary microRNA. MES, 2-(*N*-morpholino)ethanesulfonic acid; a.a., amino acid(s); MPPIX, metalloprotoporphyrin IX.

Introduction

DGCR8 is a necessary component of the microRNA biogenesis pathway [14, 46]. It also binds to the cofactor heme, forming a stable interaction with a slow off-rate while in the ferric form [28, 47]. Our lab has shown that the native DGCR8-heme complex exists in the Fe(III) form, and is extremely stable. Reduction of the heme iron in DGCR8 leads to loss of primary miRNA cleavage activity in *in vitro* assays [48]. Heme in DGCR8 is coordinated by a unique two-cysteine configuration, where cysteine-352 from both subunits of the homodimer are brought in close proximity near the dimerization sub-domain of DGCR8 (a.a. 276–350) [27]. This lead us to the thesis that heme is a stable and permanent part of the DGCR8 holoprotein, and that activation by ferric heme leads to maximal cleavage of pri-miRNA. Gong et al. have also shown that DGCR8 can be cleaved by caspases *in vitro* in the middle of the heme-binding domain, leading to loss of heme and associated loss of activity [49].

The field of porphyrin chemistry is old and expansive [50, 51, 52]. Metalloporphyrin derivatives of heme-binding proteins are made routinely [53], and have been since at least the early 1960s [54]. In addition, the chemistry involved in making any arbitrary metal-porphyrin derivative is well established, and has been useful in the past when synthesizing, for instance, antimalarial drugs. This provides us with opportunities to modify DGCR8's heme cofactor and determine the effects on DGCR8 binding and pri-miRNA processing, with the goal of understanding which features of heme are responsible for the activity of DGCR8. Here we have modified the central metal of heme, changing it from Fe to a diverse set of alternative metals. Here we show that these derivatives can have an effect on DGCR8's activity *in vitro*, and suggest some possible reasons for this.

Procedures

Fe(III)NC1, apoNC1 and apoNC1 P351A constructs were purified as described previously [47, 48]. Fe(III)PPIX and Co(III)PPIX were obtained from Sigma (St. Louis, MO, USA). All other metalloporphyrins were obtained from Frontier Scientific (Logan, UT, USA). They were dissolved in 0.1 M NaOH, and diluted to 100 µM for use in the reaction. All porphyrin solutions were stored at $\leq$ 4°C and made fresh on the day of use. The general protocol was as follows: the metalloporphyrins were dissolved and the concentrations were determined either gravimetrically or using published extinction coefficients (see Appendix Table A.1). A 100 μM porphyrin stock solution was added stepwise into a solution containing apoP351A. We looked for a spectral shift, a difference in the absorbance spectrum between the free porphyrin in solution and the porphyrin incubated with apoP351A that would suggest an interaction. To test the strength of the association, we injected the complex onto a gel filtration column to see if the protein P351A and the metalloporphyrin would co-elute, as was shown to be the case previously with P351A and ferric heme. The AKTA Purifier chromatography system allows us to monitor absorbance at three wavelengths simultaneously, allowing us to detect the elution of protein (280 nm) and two MPPIX specific absorbance peaks, one of which was always the observed Soret peak of the NC1-MPPIX 1:1 mixture.

The MPPIX-binding reaction buffer contained 1 mM DTT, 400 mM NaCl, and the buffer was either 20 mM Tris pH 8.0 (for P351A) or 50 mM MES pH 6.0 (for wild-type NC1). Absorbance spectra were collected on a Varian Cary 300 Bio spectrophotometer with bandwidth set to $\leq$ 1 nm. Reconstituted pri-miRNA processing assays were done as described previously [48], and were analyzed on 15% acrylamide, 7 M urea denaturing gels, and quantified using `ImageQuant` (GE Healthcare). The `ImageQuant` software we use allows us to draw a box around each band and quantify the total signal. For each sample box quantified, there is a Volume V_i and an Area A_i . The Volume is the total amount of signal in that box, and the Area is the size of the box itself. We draw a background box to estimate the background in our sample. The box has Volume V_B and Area A_B. We assume background is constant over the lane, meaning that doubling the area of a box will double the amount of background measured, for example. The measured Volume V_i - the raw data produced by the `ImageQuant` program - is due to the true sample volume (V_i^0) and the background (V_i^B), so $V_i = V_i^0 + V_i^B$. The true volume due to the sample alone is then $V_i^0 = V_i - V_i^B$. Since background is constant,

$$\frac{V_i^B}{A_i} = \frac{V_B}{A_B}$$

Solving for V_i^B and substituting into the equation above gives:

$$V_i^0 = V_i - A_i \frac{V_B}{A_B}$$

So, from this we can calculate the amount of signal in each band of the gel.

Results

A metal ion is necessary for association with DGCR8. In previous work [47], we have shown that a mutant version of the NC1 construct, proline 351 to alanine (P351A), can be purified as a heme-free dimer (apoP351A) when recombinantly expressed in *E. coli*. This can then be incubated with ferric heme to form a stable DGCR8-heme interaction. The wild-type NC1 can also be generated in the apo form; however, this requires several additional steps an results in a lower yield. Because of the ease of expression and handling, we chose to use association of the PPIX derivative with P351A as the first criterion for determining binding. The second criterion was whether it would likewise associate with the wild-type NC1. Incubation of Protoporphyrin IX (PPIX)

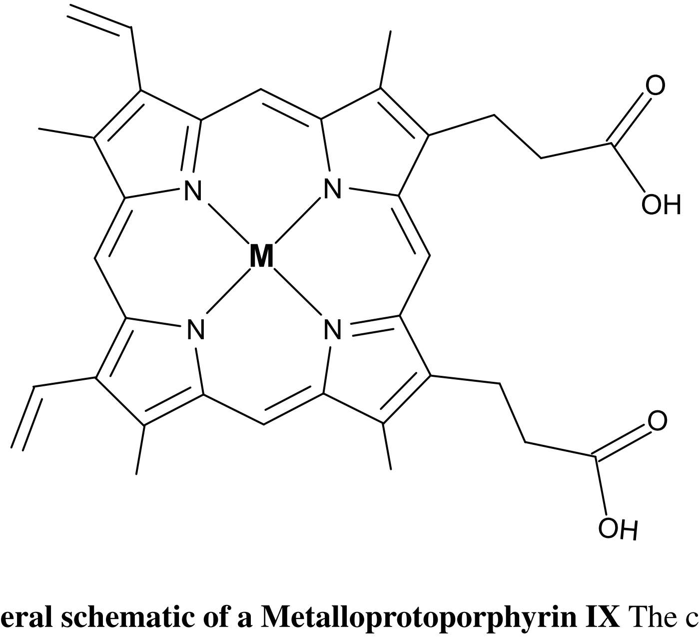

Figure 4.1: **General schematic of a Metalloprotoporphyrin IX** The central **M** stands for Fe(III), Co(III), Cr(III), Ga(III), etc.

with P351A dimer resulted in a dominant peak at 383 nm, which was slightly shifted from the 377-nm peak of PPIX without DGCR8 (Figure 4.2A). The shape of the spectra did not change as the molar ratio of PPIX and P351A dimer was increased above 1:1, indicating that the PPIX did not bind specifically to P351A. In addition, SEC analysis of the 1:1 mixture revealed no 383-nm absorbance stably associated with the P351A dimer (Figure 4.2B). We conclude based on this that a metal ion is required for the DGCR8-heme interaction.

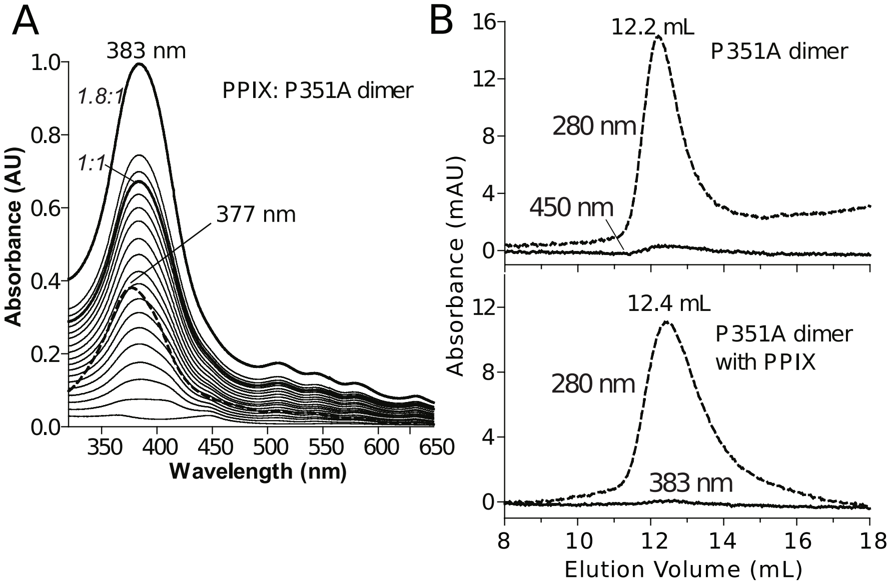

Figure 4.2: **Protoporphyrin IX (PPIX) does not associate with NC1 P351A** *A*, UV-Vis spectra of titration of PPIX into 7.5 µM heme-free P351A dimer at 0.5 µM intervals. After the 8.5 µM step, the PPIX concentration was increased directly to 13.5 µM. Some molar ratios of PPIX and P351A dimer are indicated in italics. The spectrum of free PPIX at 5.0 µM is shown in dashes. *B*, Size-exclusion chromatography analyses of 4 µM heme-free P351A dimer and its 1:1 molar mixture with PPIX shows that PPIX does not associate strongly with P351A.

Co(III) and Cr(III)PPIX associate stably with NC1 P351A and NC1 wild type. Using the screening procedure described in the previous section, we found that two of

Table 4.1: **DGCR8 Binding Properties of Metalloporphyrins.** The properties of Fe(II) and Fe(III)PPIX were known from a previous study [48] and are shown for comparison. All were tested for binding to apoP351A, and MPPIXs that showed a shift in Soret peak were tried in apoNC1 w.t.

MPPIX	Soret Peak (nm)	Shift	New Soret Peak (nm)	Co-elution	Activation
Fe(III)	380	Yes	367, 450	Yes	Yes
Fe(II)	390	No	N/A	No	No
Co(III)	392	Yes	367, 456	Yes	Yes
Cr(III)	435	Yes	437	Yes	No
Mn(III)	462	No	N/A	No	N/A
Zn(II)	413	No	N/A	No	N/A
Ga(III)	406	No	N/A	No	N/A
Ni(II)	389	No	N/A	No	N/A
Rh(III)	410	No	N/A	No	N/A
Sn(IV)	407	No	N/A	No	N/A
In(III)	409	No	N/A	No	N/A
Cu(II)	388	No	N/A	No	N/A

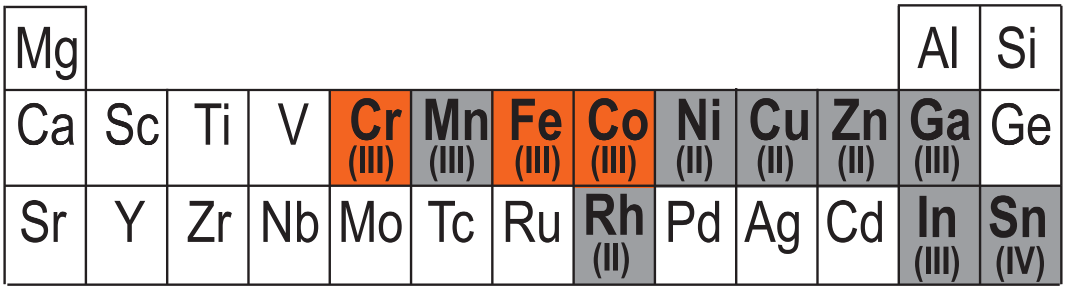

Figure 4.3: **Metalloprotoporphyrin IXs used in this study.** The protoporphyrin IX in complex with the metals shown in red bind to DGCR8 in our screen; gray were negative for binding. All were commercially available and used as received.

the 10 metalloporphyrins we tried were able to associate with P351A (Figure 4.3, Table 4.1). In the titration of Co(III)PPIX to apoNC1, we saw a hyperporphyrin spectrum for Co(III)PPIX, with peaks appearing at 368 and 456, and 563 nm, and a non-specific peak at 432 nm became prominent as the Co(III)PPIX was added in excess (Figure 4.4A). The relative heights of these peaks are similar to those seen with Fe(III)PPIX, suggesting the formation of a similar complex. The protein was then injected onto a analytical size-exclusion chromatography column, and the absorbances at 280, 432, and 456 nm were monitored. The 280 nm species co-eluted with a peak at 456 nm and a smaller one at 432 nm, indicating that the Co(III)PPIX remains bound to NC1. These peaks eluted at 12.5 mL, consistent with NC1 remaining in the dimer form after association with CoPPIX.

For Cr(III)PPIX, we saw a shift in the Soret peak from 435 to 437 nm, like that reported for a complex of Cr(III)PPIX and apomyoglobin [55]. We then moved to using the apoNC1 wild-type construct, which is more difficult to produce and handle. We saw that NC1 was able to associate with Co(III)PPIX and Cr(III) as well, showing similar spectral shifts as seen in the P351A construct (Figure 4.4).

Co(III)PPIX, but not Cr(III)PPIX, increases pri-miRNA processing activity of DGCR8. Previous work in the lab [48] showed that the heme-free NC1 construct of

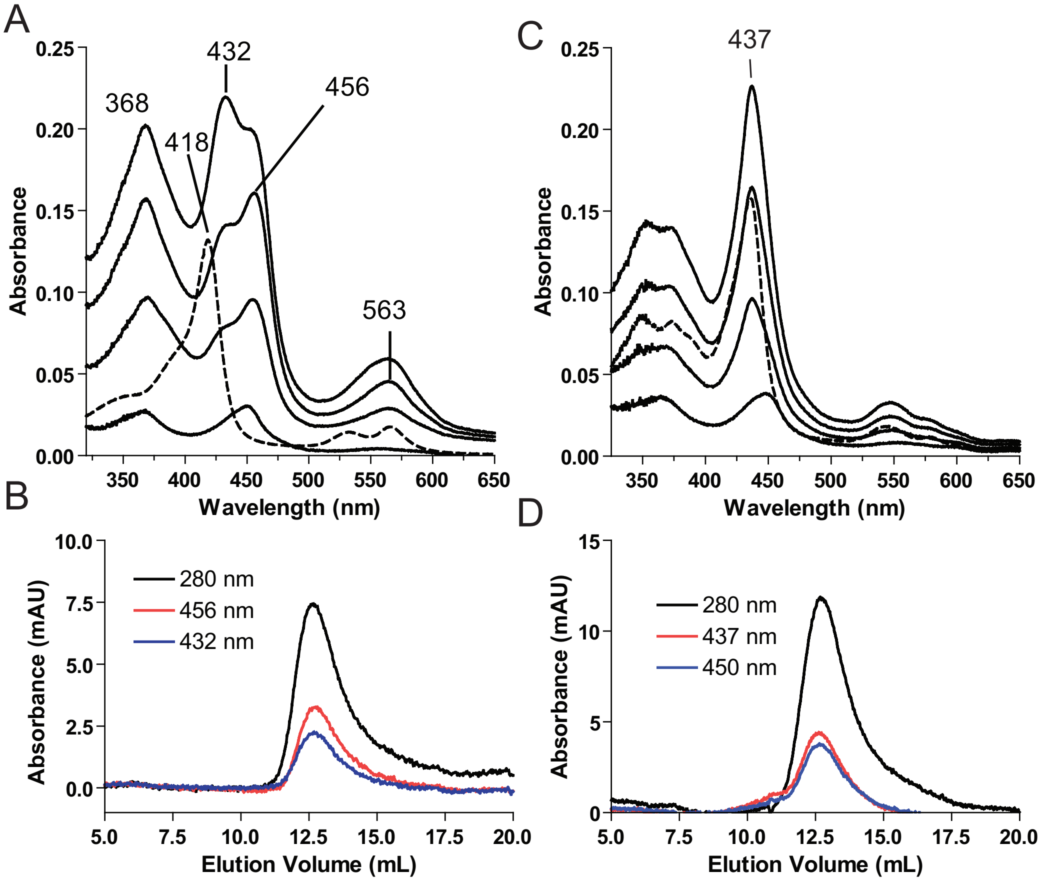

Figure 4.4: **Association of Co(III)PPIX and Cr(III)PPIX with apoNC1.** *A*, Co(III)PPIX reacted with 2 μM apoNC1 dimer. Steps show successive addition of 1, 2, 3 μM Co(III)PPIX. Addition of excess (3 μM) leads to formation of a non-specific peak at 432 nm. Absorbance spectrum of Co(III)PPIX alone is shown in dashed line. *B*, size exclusion chromatography of reconstituted Co(III)PPIX-NC1, showing the formation of a stable CoPPIX-bound dimer. *C*, Cr(III)PPIX reacted with 2 μM apoNC1 dimer. Steps show successive addition of 1, 2, 3 μM Cr(III)PPIX. Absorbance spectrum of Cr(III)PPIX alone is shown in dashed line. *D*, size exclusion chromatography of reconstituted Cr(III)PPIX-NC1, showing the formation of a stable CrPPIX-bound dimer.

DGCR8 could be activated for pri-miRNA processing by ferric but not ferrous heme. We expanded this assay to include Co(III)PPIX and Cr(III)PPIX bound NC1, in order to see the relative effects that each had on pri-miRNA cleavage by Drosha. ApoNC1 and the metalloporphyrin derivatives were combined in a 1:1 dimer:MPPIX ratio. Quantification of the cleavage product , pre-miRNA, shows that DGCR8 reconstituted with Fe(III)PPIX behaves as reported previously, increasing pri-miRNA cleavage back to native levels. Using pri-miR-380, we can see (Figure 4.5) that native NC1, as purified from *E. coli*, is active in *in vitro* processing, whereas apoNC1 is deficient. Addition of Fe(III)PPIX to apoNC1 restores this activity, as does addition of Co(III)PPIX, but not Cr(III)PPIX. Notably, Co(III)PPIX behaves consistently, activating DGCR8 to comparable degree in miR-30a and miR-380. Cr(III)PPIX is unable to increase pri-miRNA processing above.

Discussion

Metalloporphyrin derivatives of DGCR8 could allow us to up or down regulate pri-miRNA processing *in vivo*, offering research or therapeutic possibilities. It is known that Chromium and Cobalt PPIX are not degraded by heme oxygenase [56], the usual enzyme that destroys unincorporated heme in cells — in fact, they inhibit its activity [57, 58]. This has opened up possibilities for using CoPPIX and Cr(III)PPIX therapeutically to treat tumors. The situation with DGCR8 may be the same, given that we have at our disposal a PPIX derivative that activates processing (Co(III)PPIX) and one that associates with DGCR8 but doesn't activate (Cr(III)PPIX). These PPIX derivatives might qualify as therapeutics or be useful as research tools if they can be shown to increase or decrease microprocessor activity in cell culture.

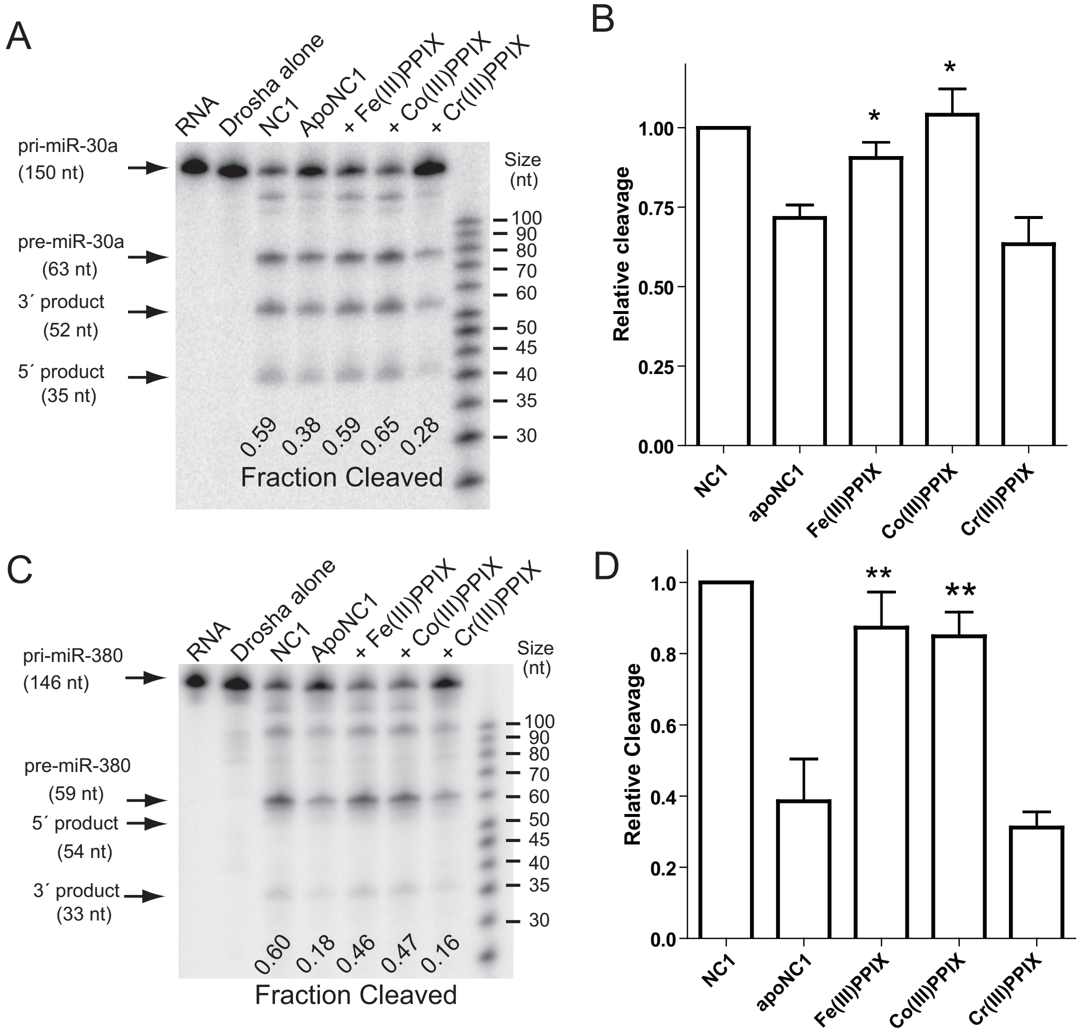

Figure 4.5: **Cleavage of ^{32}P-labeled pri-miRNA by Drosha and NC1 reconstituted with Fe(III), Co(III), and Cr(III)PPIX.** *A*, example gel of pri-miR 30a cleavage assay, with the cleavage products marked. The cleavage assay was done aerobically at 37°C, at 25 nM NC1 dimer concentration, for 30 minutes. *B*, graph of three repeats of the cleavage assay, showing the relative amount of cleavage normalized to the NC1 cleavage amount. *C* Cleavage of ^{32}P-labeled pri-miRNA 380 by Drosha and NC1 reconstituted with Fe(III), Co(III), and Cr(III)PPIX. The cleavage assay was done aerobically at 37°C, at 25 nM NC1 dimer concentration, for 45 minutes. *D* Asterisks indicates statistically significant activation of cleavage activity (** = $P \leq 0.01$, * = $P \leq 0.05$ by two-tailed Student's t-test), relative to apoNC1.

CHAPTER 5

Biophysical Characterization of DGCR8

Abbreviations Used: PPIX, protoporphyrin IX; miRNA, microRNA; DGCR8, DiGeorge Critical Region 8; pri-miRNA, primary microRNA; w.t., wild-type.

Introduction

DGCR8 and its interaction with heme have been the subject of several papers so far, two of which are reprinted here. In this chapter, I explore further the interaction of DGCR8 with heme and show unpublished results that shed light on the DGCR8-heme connection.

Procedures NC1 and NC1 P351A constructs were purified as described previously [47, 48].

Fe(III)PPIX and Co(III)PPIX were obtained from Sigma (St. Louis, MO, USA). Hemin was dissolved in 0.1 M NaOH, and diluted to 100 μM for use in the reaction. All porphyrin solutions were stored at $\leq 4°C$ and made fresh on the day of use. Circular Dicroism spectra were taken on a JASCO J-715 spectropolarimeter at 25 °C using a bandwidth of 1 nm. Stopped-flow kinetic traces were done on an Applied Photophysics SX.18MV spectrophotometer. Fluorescence data were from a Photon Technology International QuantaMaster Spectrofluorimeter.

The Fe(III)PPIX-binding reaction buffer contained 1 mM DTT, 400 mM NaCl, and the buffer was either 20 mM Tris pH 8.0 (for P351A) or 50 mM MES pH 6.0 (for wild-type NC1). Absorbance spectra were collected on a Beckman DU 530, a Beckman DU 800, or a Varian Cary 300 Bio spectrophotometer as indicated.

Results

The denaturation of DGCR8 NC1 is a multi–step process characterized by gradual loss of heme. Chaotropic agents such as urea and guanidinium chloride can be used

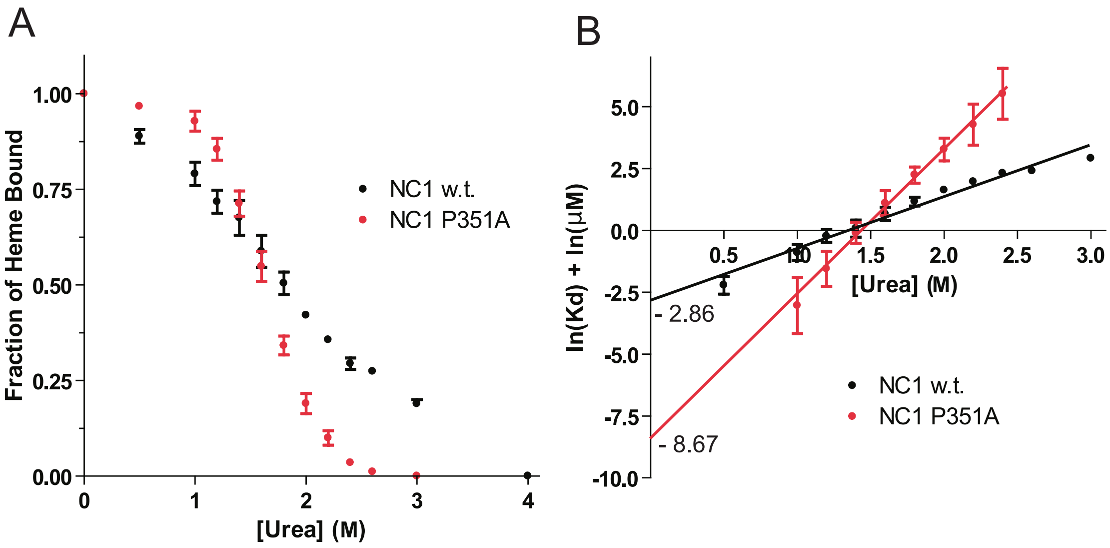

Figure 5.1: **Denaturation of heme-bound NC1 constructs in urea monitored by absorbance at 450 nm.** 8 M urea was titrated into solutions of both NC1 w.t. and NC1 P351A, and the heme absorbance at 450 nm was measured to determine fraction of heme binding. *A*, normalized absorbance readings at 450 nm over the course of the reaction. Shown are the results of three replicates, with error bars representing standard deviations. *B*, plot of the data in *A* transformed according to Equation 5.1. The *y*-intercepts are labeled.

to perturb the structure of proteins. Traditionally, we can monitor the structural changes by looking at the circular dicroism spectrum or the change in fluorescence of tryptophan or tyrosine. Since DGCR8 is a heme-binding protein with a unique absorbance peak around 450 nm, we can also monitor the loss of heme over the course of denaturation.

In order to estimate the K_D of NC1 w.t. and P351A, we can assume that, at a given urea concentration i, $K_{D,i} = [H][P]/[P \cdot H]$. We assume further that the ratio $[H]/[P \cdot H]$ is equal, at all concentrations, to the ratio of the 450 nm absorbance lost at the current urea concentration $(A_\circ - A_i)$ to the absorbance at 450 nm (A_i). This is justified due to the fact that free heme in solution has no significant absorbance at 450 nm. We can also assume that, since we start with a pure species of heme-bound DGCR8, at all times

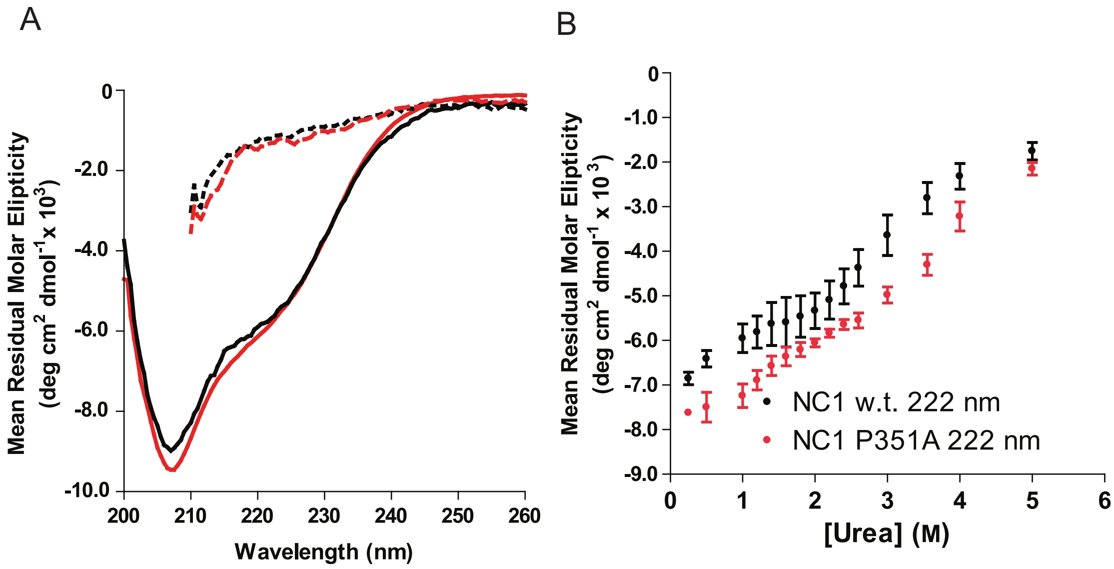

Figure 5.2: **Denaturation of heme-bound NC1 constructs in urea monitored by Circular Dicroism Spectroscopy** *A*, Native CD spectrum of the NC1 constructs. The protein concentration was 8 µM, in 20 mM Tris pH 8.0, 400 mM NaCl, 1 mM DTT. The P351A gave a similar spectrum. Dashed line is NC1 in 5 M urea, showing complete denaturation; P351A shows the same. *B*, Denaturation of heme-bound NC1 constructs in urea monitored by circular dicroism spectroscopy. 8 M Urea in 20 mM Tris pH 8.0, 400 mM NaCl, 1 mM DTT was titrated in, and the signal was monitored at 222 nm after 5 min incubation at 25°C. Signal was normalized to the protein concentration to compensate for dilution. Shown is the average of 3 determinations; error bars are standard deviations.

$[P] = [H]$. From this we can see that

$$K_{D,i} = \frac{A_\circ - A_i}{A_i}[P] = \frac{(A_\circ - A_i)}{A_i}\frac{(A_\circ - A_i)}{\varepsilon_{450}} = \frac{(A_\circ - A_i)^2}{A_i\,\varepsilon_{450}} \tag{5.1}$$

given ε_{450} is the extinction coefficient of the heme at 450 nm and $[P] = [H] = (A_\circ - A_i)/\varepsilon_{450}$. If we plot $\ln K_{D,i}$ vs. urea concentration, we can use linear regression to estimate the dissociation constant at 0 M urea, K_D. Urea is assumed to perturb the ΔG° of folding in a linear manner [59, 60, 61], so we can expect that plotting $\ln K_D$ will give us a roughly linear plot. Indeed, this is what we see in Figure 5.1B. Both the NC1 w.t and P351A proteins give us straight lines when plotted as $\ln K_D$ vs. [urea]. The significant thing is to compare the slopes, which suggest that the mechanism by which the two proteins unfold is distinct. If two proteins unfold by a similar mechanism, but have different absolute stabilities, we would expect that the data in Figure 5.1A would have the same line shape; but that one curve would be shifted to the left or right. However, what we see is that NC1 P351A gives a sigmoidal curve with a well-defined transition around 1.75 M urea, whereas NC1 w.t. has a more gradual, almost linear, folding transition. This shows on the $\ln K_D$ plot as a difference of slope, where we might expect only a difference in y-intercepts for two proteins that unfold by the same mechanism. In Figure 5.1B, we see a linear regression of the K_Ds calculated from Equation 5.1. From the y-intercept we get an estimate of the native K_D for P351A, equal to $10^{y_{int}}$ M, or roughly 2 nM. This signifies a strong interaction, though not on the level of that seen in, for example, myoglobin (10^{-14} M [62]).

To learn more about the process of heme loss, we used circular dicroism spectroscopy to monitor the unfolding of the two construct under the same conditions. Monitoring in the far UV (190–260 nm), we see a CD peak for both centered around 222 nm, characteristic of a α-helical signal (Figure 5.2). Upon addition of 5 M urea, the signal disappeared, leaving only a flat baseline (Figure 5.2A). The gradual addition of urea to the sample was followed by monitoring the CD signal at 222 nm, giving the plots seen in Figure 5.2. No unambiguous transitions can be seen in the CD signal that we can attribute directly to heme dissociation. This may be due, in part or in whole, to the fact that many of the alpha helical regions of DGCR8 NC1 are located in the C terminal region, far away from the heme binding domain. However, we do see what look like two transitions in the CD signals of both w.t. and P351A, the first of which is completed by 3 M urea. This could be correlated to the loss of heme that we see using visible spectroscopy, which is complete at around 4 M urea.

Sohn et al. showed CD data for the a.a. 493-720 region of DGCR8, showing a strong alpha-helical signal with a peak at 222 nm [26]. To the extent that the protein is modular and the domains fold and unfold independently, this means the majority of our signal could come from the RNA binding domains and the C-terminal tail. It might be worthwhile to repeat this using the HBD domain construct to see if the signal is stronger. We know that the heme binding domain contains regions of β-strands, especially the dimerization sub-domain, which is almost entirely β-strand. The CD signal for β regions is not as strong as for α, so we might need to remove a large chunk of the rest of the protein in order to better see the unfolding signal.

Stopped flow spectrophotometry shows that the rate of heme association is limited by the kinetics of heme uncomplexion. To determine the association rate of DGCR8 with Fe(III)PPIX, we assume the reaction of hemin with DGCR8 proceeds as a bimolecular interaction,

$$P + H \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} P \cdot H$$

with rate constants k_1 (forward) and k_{-1} (reverse). This model gives a dissociation constant equal to

$$K_D = \frac{[P][H]}{[P \cdot H]} = \frac{k_{-1}}{k_1}$$

The K_D has units of molarity (M). The K_D can in principle be determined from the k_{-1} and k_1 values. These rate constants give us the general rate equation,

$$\frac{\mathrm{d}[P \cdot H]}{\mathrm{d}t} = \underbrace{k_1[P][H]}_{\text{Forward}} - \underbrace{k_{-1}[P \cdot H]}_{\text{Reverse}}$$

In our association reaction, we assume a pseudo-first-order regime [64, 65]. This occurs when the concentration of one reactant is in great excess, making it effectively constant throughout the reaction. If $[P]_\circ$, the initial concentration of NC1, is much greater than $[H]$,

$$\frac{\mathrm{d}[P \cdot H]}{\mathrm{d}t} = k_1[P]_\circ[H]$$

where $[P]_\circ$ is the initial protein concentration. We have also neglected the slow reverse reaction on the basis that, for heme proteins, $k_1[P]_\circ[H] \gg k_{-1}[P \cdot H]$. This model gives

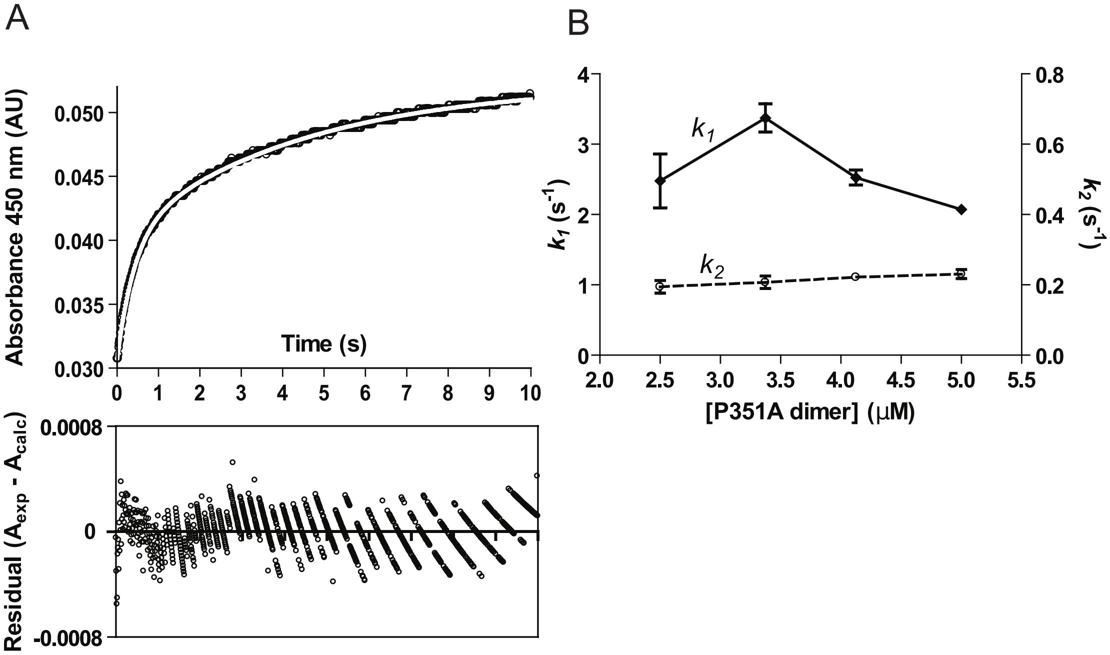

Figure 5.3: **Kinetics of the Fe(III)PPIX-DGCR8 association reaction..** *A*, a kinetic trace of 0.5 µM Fe(III)PPIX reacted with 4 µM NC1 P351A dimer, with the absorbance monitored at 450 nm. Individual data points are circles (○), and the fitted line is given in white. The data is fit to a double exponential of the form $A = B + C(1 - e^{-k_1 t}) + D(1 - e^{-k_2 t})$, where B, C, and D are constants and k_1 and k_2 are our fitted pseudo-first-order rate constants for the reaction. The lower panel is the corresponding residual plot, showing that the fit is appropriate to the data. A look at the residual plot also shows that the output of the spectrophotometer we used is subject to rounding error [63]; however, this should not effect our conclusions. *B*, the results of three technical repeats at four different NC1 P351A concentrations. Both k_1 and k_2 are roughly constant over the range of concentrations tried.

a pseudo-first-order rate constant k_1' when the reaction is fitted, with $k_1' \approx k_1[P]_\circ$.

$$\frac{\mathrm{d}[P \cdot H]}{\mathrm{d}t} = k_1'[H]$$

Substituting $[H] = ([H]_\circ - [P \cdot H])$ and rearrangement gives

$$\frac{\mathrm{d}[P \cdot H]}{([H]_\circ - [P \cdot H])} = k_1'\,\mathrm{d}t$$

Integration of both sides gives

$$\int \frac{\mathrm{d}[P \cdot H]}{([H]_\circ - [P \cdot H])} = \int k_1' \, \mathrm{d}t$$

$$\ln([H]_\circ - [P \cdot H]) = -k_1' t + C$$

Evaluation of the constant C at time $t = 0$ gives $C = \ln[H]_\circ$, so

$$\ln \frac{[H]_\circ - [P \cdot H]}{[H]_\circ} = -k_1' t$$

which rearranges to

$$[P \cdot H] = [H]_\circ (1 - e^{-k_1' t})$$

Multiplying both sides by the extinction coefficient at 450 nm gives the integrated rate equation,

$$A_t = A_{max}(1 - e^{-k_1' t}) \tag{5.2}$$

because, as $t \to \infty$, $[P \cdot H] \to [H]_\circ$. Varying the concentration of $[P]_\circ$ gives us a set of k_1' s which, when plotted as k_1' over $[P]_\circ$, should have a slope equal to k_1 that can be determined by linear regression to the equation $k_1' = k_1 [P]_\circ$.

However, these theoretical expectations appear not to hold; we see that the P351A dimer associates with heme at a roughly constant rate k_1'(Figure 5.3B); in addition, it is necessary to use a two exponential fit to determine the rate constants adequately. The second rate constant accounts for a small amount of the measured absorbance change, and signifies that there are two separate processes involved in the association reaction, of which k_2' is the lesser. We would expect, based on the model, that a linear regression of the pseudo-first-order rate constants k_1' should go roughly through the origin; however, both our fitted rate constants k_1 and k_2 are constant over different $[P]_\circ$.

One explanation for this is that the step we are measuring, the appearance of an absorbance peak at 450 nm, is not the rate-limiting step. We feel that the difficulty is due to the nature of Fe(III)PPIX in aqueous solution. It tends to form multimers in the absence of organic solvents or ligands [66, 67, 68]. The rate we measure for DGCR8-heme association may actually be an artifact of the monomer/multimer equilibrium of Fe(III)PPIX in solution. This idea has been discussed earlier by another group [69]. This means that we are currently without a method to directly measure the association of DGCR8 with Fe(III)PPIX. What we can conclude is that the actual association rate of Fe(III)PPIX with DGCR8 must be faster than the dissociation rate of the Fe(III)PPIX multimers in solution, which gives us a lower bound for the rate at around 3 x $10^3 s^{-1}$.

A Conserved Motif IPCL is required for heme binding. Earlier work has shown that mutation of the C352 in DGCR8 causes both loss of heme and loss of dimerization [28]. The P351A mutant, which was made based on the crystal structure of the dimerization domain [27], shows weaker binding to heme than the wild-type. We now know that these observations are due to the fact that C352 is the ligand for heme [47]. To investigate further, we made mutations in the region around C352 to see which residues were essential for heme binding. Table 5.1 shows the results of this mutagenesis work. It required making four additional mutation beyond the two we'd already made, P351A and C352A. We had a set of six alanine mutants in a stretch around C352. These we tested based on the following criteria:

Residue	Mutation	Heme Bound	4° Structure	Heme Removal	Heme Add
w.t.	–	Y	Dimer	N	–
S	S349A	Y	Dimer	N	–
I	I350A	N	Dimer	–	N
P	P351A	Y	Dimer	Y	Y
C	C352A	N	Monomer	–	N
L	L353A	N	Dimer	–	N
H	H354A	Y	Dimer	N	–

Table 5.1: **Determination of a Conserved motif within DGCR8's Heme Binding Domain.**

1. Did the protein purify bound to heme when expressed in *E. coli* in the presence of δ-ALA?
2. Did the protein purify in part as a Dimer, or only as a Monomer?
3. Could heme be scavenged from the protein by apomyoglobin?
4. could ferric heme be added to the protein to give a stable 450 nm peak?

From Table 5.1 we conclude that the S349A and H354A proteins were similar to the wild-type, suggesting that mutations had little effect on heme association. The others show an effect on heme binding that ranges from slight perturbation (P351A) to complete loss of heme association (I350A, L353A). It is not yet known whether I350 and L353 can be substituted with other amino acids to support high-affinity interaction with heme. S349A was expressed only as a dimer and was fully occupied by heme, resembling the wild type. Interestingly, all the residues required for heme-binding in this region, plus Pro351, are completely conserved among the DGCR8 family members, whereas Ser349 and His354, non-essential for heme binding, are partially conserved. The correlation between heme binding and sequence conservation indicates the importance of heme-binding to the function of DGCR8.

Heme Binding to NC1 P351A Quenches Trp/Tyr Fluorescence. Fluorescence spectroscopy of proteins allows us to probe conformational changes and interactions between optically active molecules within a protein [41]. The interaction of heme with DGCR8 was expected to correlate to a conformational shift which would show as a change in the fluorescence signal upon heme binding. We take advantage of the native fluorophores present in proteins, most significantly tryptophan and tyrosine. Tryptophan has a strong excitation band in the 275-295 range, and it is this property which we are taking advantage of in this case. We found that association of DGCR8 with heme will cause a change in fluorescence signal, and we are able to draw the following conclusions from this data. Figure 5.4A show a set of fluorescence spectra of the

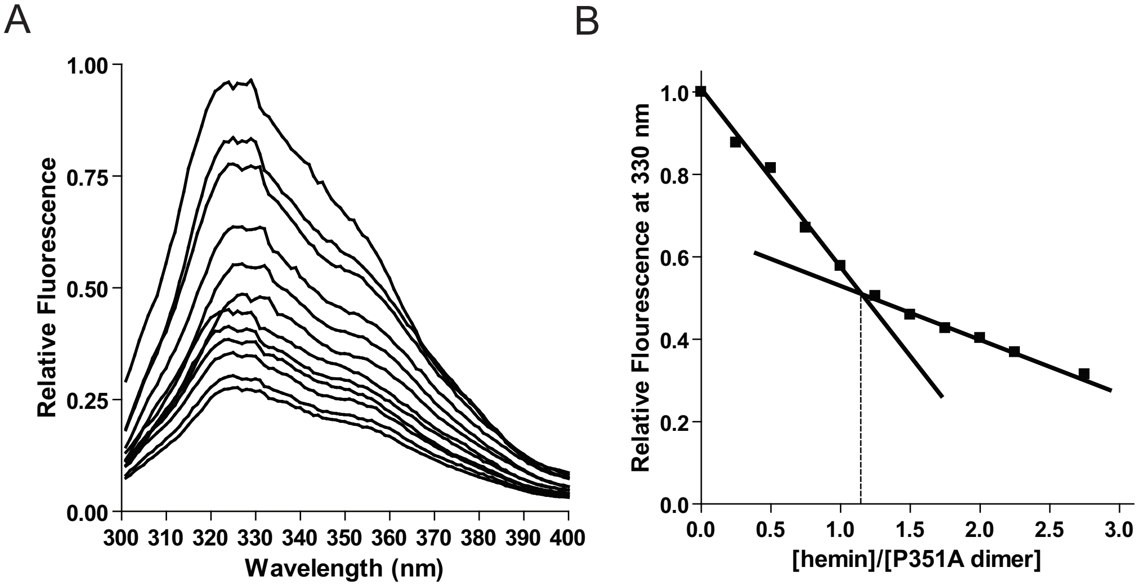

Figure 5.4: **Heme association with P351A monitored by changes in Trp/Tyr Fluorescence.** 2 µM apoNC1 P351A was titrated with ferric heme at pH 8.0, 400 mM NaCl, 1 mM DTT. Heme was added in 0.5 µM steps until greater than 2:1 excess of heme was achieved.

NC1 P351A apoprotein titrated with Fe(III)PPIX. A gradual loss of the peak centered at around 330 nm follows the addition of Fe(III)PPIX to the solution. This is partially due to the fact that heme absorbs at 290 nm as well, so it contributes to the decrease in signal seen here. However, looking at just the relative fluorescence values for 330 nm, we can see that there is a point of inflection at around 1:1 dimer:Fe(III)PPIX ratio, suggesting that there are two processes occurring simultaneously. The first is a slow, linear loss of signal due to absorbance of Fe(III)PPIX at 290 nm. The second is another, unknown process in which Fe(III)PPIX associates specifically with P351A and causes a conformational change or fluorescence quenching in one or more of the fluorophores present in the protein. NC1 P351A contains several possible residues; in the heme binding domain dimer there are 6 tryptophans and 8 tyrosines which could possibly be perturbed during the conformational shift that accompanies heme binding.

This data shows that the titration of Fe(III)PPIX into apoNC1 P351A has a corresponding conformational change, one that is saturable upon reaching 1:1 stoichiometry. This is characteristic of a specific interaction between P351A and hemin.

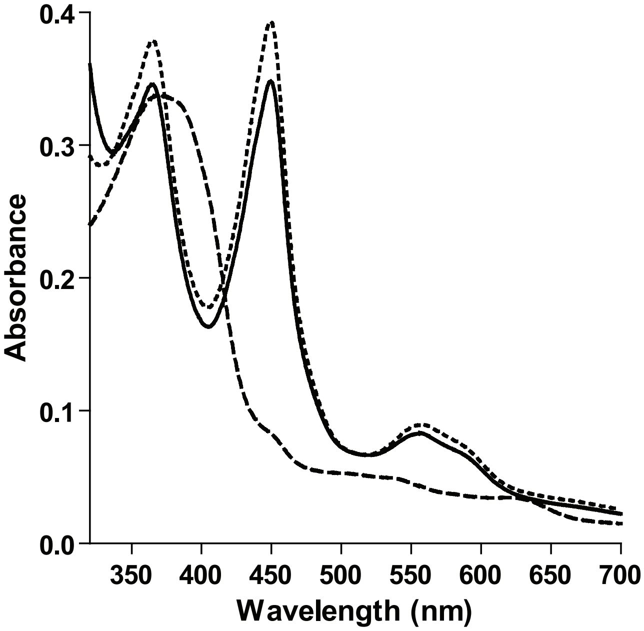

Figure 5.5: **Fe(III)PPIX-DGCR8 association is reversibly disrupted by mercury** Dotted line: 8 µM Human HBD C430S before incubation with 10 µM methylmercury acetate. Dashed line: After incubation. Solid line: after addition of 50 mM β-mercaptoethanol and 1 min at room temperature.

Removal of heme-binding by incubation with mercury is reversible. In a previous work [47] we showed that the interaction of heme with DGCR8 could be disrupted by the addition of mercury ions, which displace the thiol groups normally coordinating the heme iron to the protein. Here we show that this process is reversible. In Figure 5.5, the Human HBD construct, which contains only one cysteine per subunit, is incubated with a molar excess of MeHg(II). The heme dissociation is shown by the loss of the 366 and 450 nm absorbance peaks, which are replaced with one similar to those seen in free heme. Upon the addition of 70 mM β-mercaptoethanol, a thiol-containing reducing agent, the 450 nm and 366 nm peaks reappear within a minute. This can be explained by assuming that the mercury complexed to the DGCR8 has been competed away by the excess thiols, and the heme is free to reattach to the cys 352s of the HBD.

The fact that heme binding is disrupted upon addition of mercury, and binds again upon addition of excess thiols, suggests that the heme-DGCR8 interaction is largely stabilized by the complex of Fe(III) and cysteine. Previous work, reported in **Chapter 2**, shows that heme, once displaced by mercury, can easily be scavenged by apomyoglobin in a reaction that occurs in less than a minute. From this we conclude that the heme is loosely associated with DGCR8, or perhaps even free in solution.

CHAPTER 6

Evidence of a novel RNA secondary structure in the coding region of HIV-1 *pol* gene

(*RNA* 14, 2478-2488.)

Evidence of a novel RNA secondary structure in the coding region of HIV-1 *pol* gene

QI WANG,[1,2] IAN BARR,[1,3] FENG GUO,[1,3] and CHRISTOPHER LEE[1,2]

[1]Molecular Biology Institute, University of California at Los Angeles, Los Angeles, California 90095, USA
[2]Center for Computational Biology, Department of Chemistry and Biochemistry, Institute of Genomics and Proteomics, University of California at Los Angeles, Los Angeles, California 90095, USA
[3]Department of Biological Chemistry, David Geffen School of Medicine, University of California at Los Angeles, Los Angeles, California 90095, USA

ABSTRACT

RNA secondary structures play several important roles in the human immunodeficiency virus (HIV) life cycle. To assess whether RNA secondary structure might affect the function of the HIV protease and reverse transcriptase genes, which are the main targets of anti-HIV drugs, we applied a series of different computational approaches to detect RNA secondary structures, including thermodynamic RNA folding predictions, synonymous variability analysis, and covariance analysis. Each method independently revealed strong evidence of a novel RNA secondary structure at the junction of the protease and reverse transcriptase genes, consisting of a 107-nucleotide region containing three stems, A, B, and C. First, RNA folding calculations by mfold and RNAfold both predicted the secondary structure with high confidence. Moreover, the same structure was predicted in a diverse set of reference sequences in HIV-1 group M, indicating that it is conserved across this group. Second, the predicted base-pairing regions displayed markedly reduced synonymous variation (approximately threefold lower than average) in a data set of 20,000 HIV-1 subtype B sequences from clinical samples. Third, independent analysis of covariation between synonymous mutations in this data set identified 10 covariant mutation pairs forming two diagonals that corresponded exactly to the sites predicted to base-pair in stems A and B. Finally, this structure was validated experimentally using selective 2′-hydroxyl acylation and primer extension (SHAPE). Discovery of this novel secondary structure suggests many directions for further functional investigation.

Keywords: HIV-1 *pol* gene; RNA secondary structure; thermodynamic prediction; covariation; synonymous variability; SHAPE

INTRODUCTION

HIV is the causative agent of AIDS, now a worldwide epidemic. One serious problem for the treatment of AIDS is HIV's ability to rapidly develop resistance to antiretroviral drugs. The majority of FDA-approved anti-HIV drugs target the protease and the reverse transcriptase in the HIV *pol* gene (Simon et al. 2006). In order to better understand the development of drug resistance, it may be important to understand the structure and function not only of the protease and reverse transcriptase proteins, but also of the *pol* gene itself, such as possible RNA secondary structures, since these could affect its function.

A number of RNA secondary structures have been identified in different parts of the HIV genome (Paillart et al. 2002; Abbink and Berkhout 2003; Damgaard et al. 2004; Hofacker et al. 2004; Ooms et al. 2007). There are some well-studied examples, such as the *trans*-activating responsive (TAR) element at the 5′-end of the genome (Berkhout 1992), the Rev response element (RRE) in the *env* gene (Malim et al. 1989), and the *gag*/*pol* frame-shift hairpin (Parkin et al. 1992). They all have been found to play important roles in HIV transcription. In addition, it has been reported that an RNA secondary structure in the *gp120* gene facilitated recombination, creating a recombination hot spot in HIV (Moumen et al. 2001; Galetto et al. 2004). All these studies suggest that RNA secondary structure in HIV plays important roles in the viral life cycle. One study has suggested a relationship between RNA secondary structure and drug resistance mutations in HIV (Schinazi et al. 1994).

Thus, one important goal is the complete identification of all RNA secondary structures in HIV, particularly in regions involved in drug resistance. This requires several different kinds of analysis. Energy-based RNA folding prediction programs can give useful predictions of likely structures, but are not in and of themselves adequate evidence for a specific structure. Comparative genomic methods provide a variety of ways to test such predictions (Mathews and Turner 2006). First, comparison of many related sequences can evaluate whether regions containing predicted secondary structures are more strongly conserved than neighboring regions. Furthermore, by focusing such analysis on synonymous sites, it is possible to distinguish whether conservation is due to selection pressure on the amino acid sequence (i.e., protein function) or on the RNA sequence itself (consistent with a functionally important RNA secondary structure). Second, comparative genomics can evaluate whether the predicted secondary structure is conserved over a broader evolutionary clade. Finally, if sufficient data are available, mutation covariance analysis can directly indicate pairs of nucleotides that appear to be base-paired by identifying compensatory mutations. All of these approaches depend on having enough sequences to obtain statistically significant results.

The combination of energy-based folding and comparative genomic approaches has successfully detected RNA secondary structures in HIV. Hofacker et al. (1998) correctly identified the two well-known secondary structures TAR and RRE via a combination of thermodynamic structure prediction with phylogenetic comparison of as few as 13 full genomic sequences. The emergence of larger HIV sequence data sets provides a useful opportunity to take greater advantage of comparative genomics to identify all RNA secondary structures in HIV. Peleg et al. (2002) applied a combination of secondary structure prediction and the conservation assessment method to 178 *env* sequences and identified one novel RNA structure in the *env* gene. They also predicted another secondary structure in the *nef* gene through analyzing 106 *nef* sequences (Peleg et al. 2003).

In this study, we have applied thermodynamic prediction methods along with comparative sequence analyses of about 20,000 *pol* gene sequences, yielding strong evidence of a novel RNA secondary structure at the junction between the protease and reverse transcriptase coding regions.

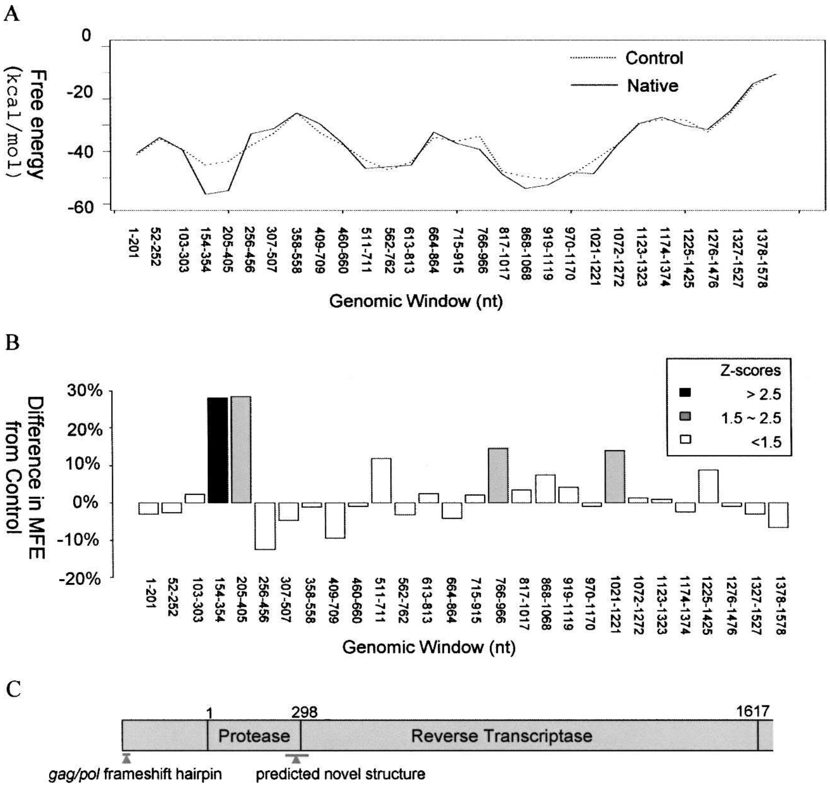

FIGURE 1. Minimal free energy (MFE) of 201-base fragments in the *pol* gene and its comparison with that of shuffled control sequences. (*A*) The *x*-axis is genomic coordinates for the 201-base fragments with respect to the first nucleotide in the protease; the *y*-axis is the MFE of the native fragment (solid line) and the mean MFE of 1000 corresponding shuffled control sequences (dotted line). (*B*) The *x*-axis is the same as in *A*; the *y*-axis is the difference in MFE of the native fragment from the mean value of 1000 shuffled control sequences. Shading indicates the significance of such difference in *Z*-score, which is the number of standard deviations from the mean of the shuffled control sequences. (*C*) Schematic diagram of the first half of the *pol* gene, up to the end of the reverse transcriptase. The genes and the positions of RNA secondary structures are marked. Coordinates are given starting from the first nucleotide in the protease, as in the rest of the article.

RESULTS

Identification of a 107-nucleotide predicted RNA secondary structure

To search for possible RNA secondary structures in the *pol* gene, we ran RNA folding predictions for 201-nucleotide (nt) windows with a step size of 51 nt throughout the HIV-1 group M consensus gene sequence (starting from the first nucleotide of protease; see Materials and Methods), using the program RNAfold (Fig. 1; see Materials and Methods for details). To assess the significance of predictions in each window, we compared the minimal free energy (MFE) of the consensus gene sequence versus that of shuffled control sequences. Specifically, we calculated the Z-score for this difference (Fig. 1B; Simmonds et al. 2004) by generating a sample of 1000 shuffled sequences using DicodonShuffle (Katz and Burge 2003; see Materials and Methods), which preserves the encoded protein sequence, codon usage, and dinucleotide composition of the original sequence.

These analyses identified a region from nucleotides 205 to 354 that yielded a predicted RNA secondary structure (Fig. 2) with statistical significance. The same stem–loop structure was predicted in two adjacent 201-nt windows (nucleotides 154–354 and 205–405) with Z-scores (from the sequence shuffling tests) of 2.7 and 2.3, respectively. These Z-scores were statistically significant, yielding P-values of 0.003 and 0.012, respectively. Other predictions with weaker Z-scores were found between nucleotides 766 and 966 ($Z = 2.1$) and 1021 and 1221 ($Z = 2.1$).

A new set of predictions using the program mfold (Zuker 1989) was generated for nucleotide windows 154–354 and 205–405. In both windows, mfold predicted an identical secondary structure of 107 nt formed by nucleotide fragments 217–323 (Fig. 2), the same structure as predicted by the RNAfold. This structure, located at the junction of the protease and the reverse transcriptase (Fig. 1C), consisted of three base-paired stems (A, B, and C) and two loops, one on the end of stem B and the other on the end of stem C. Stem A, the longest, contained 19 base pairs (bp) of complementary sequence with only a single mismatch (at nucleotide positions 225/315). Stem B consisted of 9 bp of complementary sequence, separated by an 11-nt hairpin turn. Stem C had only 6 bp of complementary sequence.

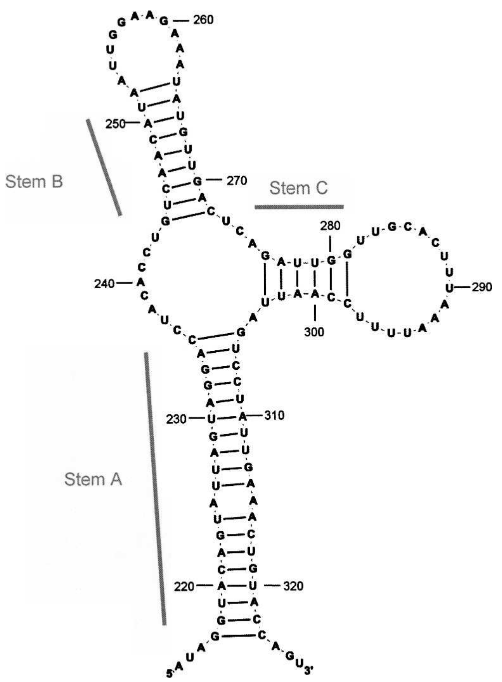

FIGURE 2. The predicted RNA secondary structure of nucleotide region 214–326 (with respect to the first nucleotide of the protease) of HIV-1 group M consensus sequence, using RNAfold with default parameters. Only base-pairings with probabilities >0.5 are shown. Stems A, B, and C are indicated.

Since the accuracy of structure prediction can be significantly improved by analysis of multiple sequences (Mathews and Turner 2006), we tested whether the predicted secondary structure is conserved over a diverse family of related HIV types. We generated secondary structure predictions for each member of the set of reference sequences representing strains of HIV group M from the Los Alamos HIV Sequence Database (see Materials and Methods). This data set consisted of a total of 37 reference sequences representing 11 distinct HIV subtypes, including subtype B. There were about four sequences for each subtype, broadly representing the subtype. For each reference, we extracted nucleotide regions 214–326 (with reference to the group M consensus in the multiple sequence alignment) and predicted the secondary structures using mfold (Zuker 1989) with default parameters. Stem A (Fig. 2) was conserved in 100% of the predicted structures; stem B was conserved in 70% of the predicted structures; stem C was conserved in 50% of the predicted structures (see Supplemental Table 1 for details). In some cases, mfold predicted multiple structures with similar energies. Overall, 100% of the reference sequences had both stems A and B in at least one of the predicted structures and 92% of the reference sequences had stem C in at least one of the predicted structures. This suggests that the predicted structure is highly conserved among different subtypes in group M. Among the three stems in the predicted structure, stem A is the most conserved, whereas stem C is the least.

We also analyzed this region using RNAz (Gruber et al. 2007), a program for predicting structurally conserved and thermodynamically stable RNA secondary structures in multiple sequence alignments. The multiple sequence alignment of the 37 group M reference sequences (see Methods and Materials) was analyzed using the program's default parameters except for a window size of 120 nt and a step size of 10 nt. RNAz predicted the same stem–loop secondary structure with the probability of 1.00. This high level of confidence suggests that the patterns of sequence conservation in HIV-1 group M are strongly consistent with this predicted RNA secondary structure.

Validation by synonymous polymorphism data from 20,000 HIV-1 sequences

The region containing the predicted RNA secondary structure is part of an open reading frame encoding the protease and reverse transcriptase proteins. Thus, conservation of nucleotide sequences can simply result from the conservation of protein sequences. Therefore, it is important to ask whether conservation in this region might simply be due to amino acid selection pressure (i.e., selection on the protein sequence), as opposed to conservation of the proposed RNA structure. Since synonymous mutations do not alter the amino acid sequence and thus are unaffected by amino acid selection pressure, measurements of conservation using only synonymous substitutions (such as ds, Ks) can distinguish specific evidence of selection on RNA structure and function from the more typical pattern of amino acid selection pressure (Xing and Lee 2006). The obvious question is whether the predicted base-pairing regions (i.e., stems A, B, and C) show markedly lower synonymous substitution rates than surrounding regions of the gene. Very large data sets of clinical HIV-1 sequences make it possible to answer this question by measuring synonymous variability for each codon position.

We computed a measure of synonymous variability from a data set of about 20,000 HIV-1 subtype B sequences (Chen et al. 2004). Following the previous analysis of Gog et al. (2007), synonymous variability was calculated for each codon position as the normalized mean pairwise distance (MPD) (raw data shown in Supplemental Fig. 1; see Materials and Methods for details). To ensure sufficient diversity of the data and minimize potential bias from phylogenetic effects, we had excluded sequences with <2% nucleotide diversity, so that in the final set including 20,000 sequences any two sequences were >2% different from each other. Over the 450 codon positions analyzed in the *pol* gene, the average MPD score was 0.206.

In order to reveal regions with suppressed synonymous variability, we calculated the moving average of the MPD score over a sliding window of 10 codons (Gog et al. 2007). The windowing analysis revealed five regions in the *pol* gene with unusually low synonymous substitution rates (Fig. 3A, I–V). Strikingly, regions IV (MPD value of 0.09) and V (MPD value of 0.07) corresponded directly to the complementary sequences of stem A in the predicted RNA secondary structure (Fig. 2). By contrast, the low synonymous variability of region I resulted from the fact that the first 13 codons of protease overlap with the reading frame of *gag*, the previous gene, and thus synonymous sites in this region of the *pol* gene were actually nonsynonymous sites (which, typically, have much lower variability, due to amino acid selection pressure) in the *gag* gene. We have not investigated the possible causes of region II or III.

We next examined in detail the correspondence between synonymous variability of individual sites and the predicted stem–loop structure (Fig. 3B). Both halves of stem A displayed unusually low levels of synonymous variability (MPD values of 0.02–0.10). Strikingly, the one mismatch in the predicted stem A structure, formed by positions 225 and 315, had dramatically higher levels of synonymous variability (MPD 0.32 and 0.18, respectively, similar to the average level of MPD in *pol*) than the predicted base-paired positions in stem A (mean MPD 0.04 with standard deviation 0.02). We evaluated the statistical significance of the low synonymous variability of the positions predicted to base-pair in stem A using the Wilcoxon rank test. This result was strongly significant (*P*-value of 3×10^{-4}). Stem B also appeared to have reduced synonymous variability, although this result was weaker (*P*-value of 0.059), consistent with the short length of stem B, and stem C showed no significant reduction (*P*-value of 0.90). We also analyzed the synonymous variability of the two weaker predictions of base-paired stems between nucleotides 766–966 and 1021–1221; neither was statistically significant.

HIV-1 has been shown to have a tendency to become A-rich (van Hemert and Berkhout 1995). To assess the potential bias resulting from this nucleotide sequence pressure, we have calculated the percentage of A in the 107-nt region of the *pol* gene that folds into the secondary structure and compared it with that of the entire *pol* gene. The percentage of A is 32% (34/107) for this region and 39% (1190/3015) for the full *pol* gene, showing no

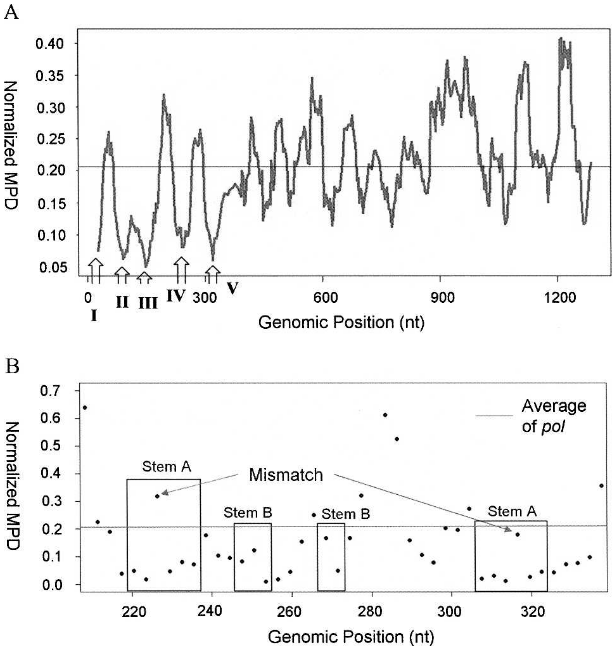

FIGURE 3. Synonymous variability in the coding region of the *pol* gene. (*A*) The moving average of synonymous variability in windows of 10 codons with a step size of one codon. The variability of each codon is measured by normalized mean pairwise distance (MPD). The *x*-axis is average genomic coordinates of codons in the scanning window with respect to the first nucleotide in the protease; the *y*-axis is the averaged MPD score in the window. The five low synonymous variability regions, regions I–V, are indicated by arrows. The horizontal line indicates the average MPD score of the *pol* region. (*B*) The synonymous variability of the codon positions in the sequence region that folds into the predicted structure. The *x*-axis is the genomic coordinates as in *A*; the *y*-axis is the MPD score. The horizontal line indicates the average MPD score of the *pol* region. The base-pairing regions of stems A and B are indicated by boxes.

significant difference (Fisher's exact test P-value = 0.3). Thus the data do not support A-pressure as an explanation for the low mutation rate observed in the secondary structure region.

Validation of specific base pairs by analysis of covariant substitutions

With extremely large sequence data sets, it becomes possible to search for validation of base pairing by finding pairs of mutations that show strong covariance, indicating that they are "compensatory mutations" (Chen et al. 1999). Again, in a protein-coding region such as *pol* it is extremely important to screen out patterns of covariance that might be due to amino acid selection pressure (i.e., selection for compensatory mutations due to their effects on protein function) (Wang and Lee 2007). Since synonymous mutations do not alter the amino acid sequence, covariance between synonymous mutations cannot be due to amino acid selection pressure. We therefore analyzed covariance between all possible pairs of synonymous mutations in *pol*, as an independent test of possible base pairings, using the same data set of 20,000 HIV-1 subtype B sequences (Chen et al. 2004).

This analysis highlights one striking feature of the predicted RNA secondary structure (Fig. 4A). By random chance, there is only a 1/3 probability that synonymous sites in one strand of an RNA stem will be base-paired to synonymous sites in the other strand. From this point of view, it is striking that synonymous sites base-paired only with synonymous sites in all three stems in the predicted RNA secondary structure (P-value = 0.037).

Covariance analysis yielded a sparse scatter of pairs of sites with statistically significant covariation (Fig. 4B; complete data in Supplemental Fig. 2; see Materials and Methods for details). No pattern was evident except for two clusters of strongly covarying pairs, each forming a line perpendicular to the diagonal, indicative of the antiparallel orientation required for a base-pairing hairpin structure. These two clusters corresponded directly to positions

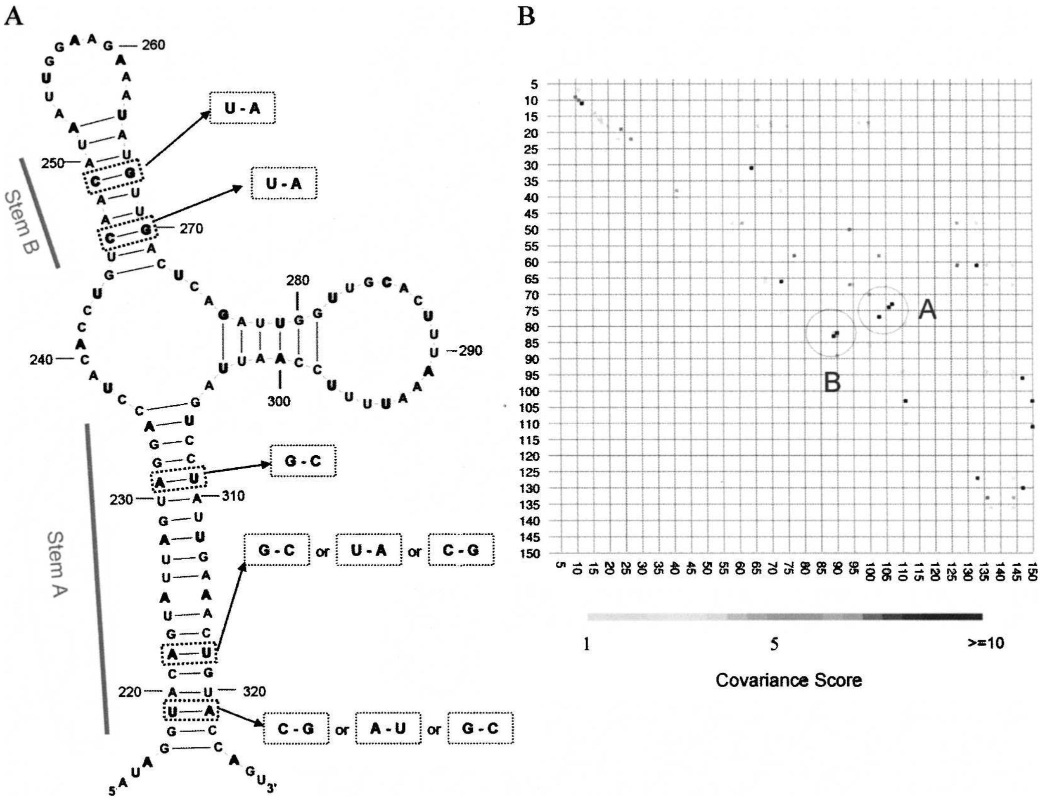

FIGURE 4. Covariant synonymous mutations in the predicted RNA secondary structure. (*A*) The predicted structure of nucleotide regions 214–326 (with respect to the first nucleotide of protease) of HIV-1 group M consensus sequence, using RNAfold with default parameters. Only base-pairings with probabilities >0.5 are shown. The covariant substitutions are given in the boxes for each pair of covariant sites. The third nucleotide in each codon is in bold. Stems A and B are indicated. (*B*) The covariation map shows the strength of covariance of synonymous substitutions for each codon position pair. The *x*- and *y*-axes are coordinates for the codon positions, starting from the first codon in the protease. For a given codon position pair, the highest covariance value for any pair of synonymous mutations at the two positions is displayed on the map. The two clusters of strongly covarying sites are highlighted in circles and labeled with the corresponding stem index A or B.

predicted to be base-paired in stems A and B, respectively. Not only did the covarying pairs match precisely the positions predicted to pair, but 9 out of 10 pairs of nucleotide mutations formed standard Watson–Crick base pairs (see Fig. 4A and Supplemental Table 2). In addition, two pairs of sites in stem A each contained three pairs of covariant substitutions (Table 1). These covariance results were highly statistically significant, by both the chi-square test and Fisher's exact test (P-value $< 10^{-10}$; see Materials and Methods). The existence of multiple covariant sites in stems A and B provides compelling independent evidence for the precise base pairs predicted by the RNA folding calculations.

To assess the significance of the synonymous covariation results as evidence for the predicted secondary structure, we performed several statistical tests. Considering all possible pairs of synonymous mutations within the 1.4-kb region analyzed in this study, there are a total of 124,663 possible synonymous mutation pairs that would satisfy Watson–Crick base pairing. Thus, to obtain a significance level of 0.01 after the Bonferroni correction, we set a P-value cutoff of 10^{-7} for any specific synonymous mutation pair. Only 36 pairs (0.03% of the total) met this criterion for significant covariation. By contrast, of the only 16 possible pairs of complementary synonymous mutations at the pairing sites in the three stems of the proposed secondary structure (Fig. 4A), nine (or 56% of the total) were actually detected by this criterion as showing significant covariation. Indeed, the P-values for these pairs were strong (P-value ranging from 10^{-8} to 10^{-42}). These data provide very strong validation for the predicted secondary structure, because they do not merely confirm the general region of the predicted stem, but specifically validate the precise base-pairing of the predicted secondary structure. This is a statistically strong result. The likelihood of obtaining these results by random chance is only 5.4×10^{-29} (see hypergeometric test, Materials and Methods).

Covariant substitutions were observed not only in HIV-1 subtype B sequences but also in subtype H. In the group M consensus sequence, there was one paired C and G at nucleotide positions 249 and 267, respectively, which was located in stem B. This pair of sites remained complementary after substitutions at both positions in subtype H, to nucleotides U and A, respectively. The covariant substitutions provide independent confirmation of the secondary structure inferred from both the synonymous variability analysis and the free energy-based folding.

Experimental validation using the SHAPE assay

We probed the secondary structure using selective 2′-hydroxyl acylation and primer extension (SHAPE). SHAPE exploits the difference in reactivity between paired and unpaired nucleotides toward *N*-methylisatoic anhydride (NMIA) (Merino et al. 2005; Wilkinson et al. 2006). NMIA modifies an RNA at a 2′ hydroxyl, and the modification blocks subsequent primer extension using a reverse transcriptase. The products of reverse transcription are examined using sequencing gels, enabling direct readout of the modification pattern. Increased gel band intensities in the NMIA-treated lane (+) versus the untreated lane (−) identify RNA residues that are reactive and thus likely unpaired.

A 113-nt sequence fragment of a HIV-1 subtype A1 reference (see Materials and Methods) that contains the proposed structure was synthesized using in vitro transcription and was subjected to SHAPE analysis (Fig. 5). We observed high reactivity (>50%) for the nucleotides in the loops of the B and C stems, the junctions between stems A and B and stems C and A, as well as the 5′ and 3′ flanking regions. In contrast, stems A and B show little NMIA reactivity, indicating that they are strongly base-paired as in the predicted secondary structure. Stem C, the shortest

TABLE 1. Evidence of covariation for pairs of nucleotide sites

		Nucleotide 219			
	Observed	A	G	C	U (wt)
Nucleotide 321	G	0	0	**50**	166
	U	**4**	1	0	3
	C	0	**4**	0	3
	A (wt)	42	76	93	20,584
		Nucleotide 222			
	Observed	G	U	C	A (wt)
Nucleotide 318	A	0	**7**	0	50
	G	0	0	**7**	56
	C	**12**	0	0	26
	U (wt)	54	6	13	20,798

The numbers indicating the standard Watson–Crick base pairs are shown in bold.
(wt) the consensus.

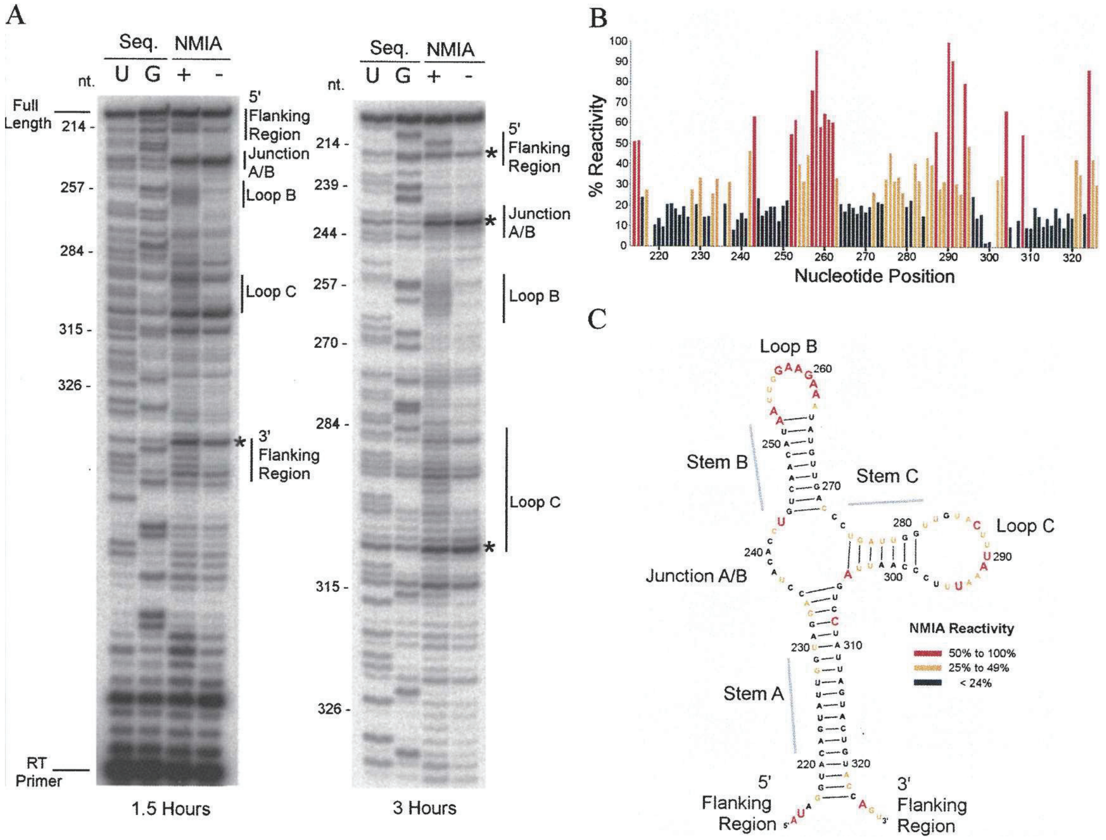

FIGURE 5. SHAPE analysis of the proposed HIV secondary structure. (*A*) Sequencing polyacrylamide gels showing SHAPE analysis of the RNA after running 1.5 h (*left*) and 3 h (*right*) at 60 W, as was necessary to resolve bands on the 5′ and 3′ ends. The lanes marked U and G are sequencing lanes, in which ddATP and ddCTP were added to the reverse transcription to cause chain termination at U and G sites. NMIA lanes (+) and (−) correspond to incubation with 60 mM NMIA for 45 min at 37°C and a control with NMIA omitted. The flexible segments are visible as regions of increased NMIA modification in the (+) lane as compared to the (−) lane. The asterisks indicate the pauses of reverse transcription even in the absence of NMIA, likely caused by the helices A and C. (*B*) SHAPE modification intensities, determined from the gels in *A*. SHAPE intensities of residues 214–309 were determined from the 3-h gel, and 310–326 from the 1.5 h gel. Bars show the amounts of modification at each position relative to the most highly modified nucleotide. (*C*) SHAPE modification intensities mapped onto the secondary structure, showing regions of heavy modification at unpaired nucleotides. The structure was generated using RNAfold. Only base pairings with probabilities >0.5 are shown. Residues with intensities between 25% and 49% and 50% and 100% are shown in orange and red, respectively. The sequence used in this analysis was retrieved from a reference genome of HIV-1 subtype A1 (GenBank accession number: U51190) from the Los Alamos HIV Sequence Database.

helix, shows modest to low degrees of modification. In addition, the reverse transcriptase tends to pause at the 3′ end of strong secondary structures. Under the conditions we used, strong pauses in reverse transcription were observed for stems A and C even in the absence of NMIA (Fig. 5A, asterisks), consistent with our predicted secondary structure. Therefore, the SHAPE data support the novel 3-stem RNA structure that we have proposed.

DISCUSSION

Detection of an RNA secondary structure using evolutionary data

The first indication of this novel secondary structure came from our previous studies of HIV mutation covariance in clinical HIV samples (Chen et al. 2004; Wang and Lee 2007), which identified the unusual pattern of synonymous mutation covariance shown in Figure 4. The clear ‘‘diagonal’’ pattern in the covariance matrix, the remarkable conservation of Watson–Crick base-pairing by compensatory mutations, and the obvious sequence complementarity of these regions immediately indicated an RNA secondary structure.

Subsequent secondary structure prediction using both thermodynamic prediction methods as well as methods that exploit the large amount of comparative sequence data—around 20,000 clinical HIV samples—yielded a consistent folding prediction across the HIV-1 group M family. The thermodynamic prediction programs mfold and RNAfold both predicted a secondary structure matching the base pairing seen in our covariance data. In addition, the folding free energy of the native sequence was significantly lower than those of shuffled control sequences (Fig. 1), indicating a statistically significant result. Moreover, stems A and B persisted in the predictions of all the sequences throughout the diverse set of reference sequences of HIV-1 group M. Also, the base-paired regions of the predicted structure were highly conserved at the nucleotide level, even at synonymous sites, which have no effect on the protein sequence (Fig. 3). In the coding regions of viral genomes, reduced synonymous mutation rates have been found to be a signature for nucleotide secondary structures (Simmonds and Smith 1999) or other nucleotide regulatory sequences (Gog et al. 2007). It should be noted that our bioinformatics analysis focuses on positive-strand RNA, because the HIV life cycle includes only positive-strand RNA, and the biological relevance of the negative strand is not clear.

It should be emphasized that these evolutionary data provide evidence not only of a secondary structure, but also of its functional importance to the reproductive fitness of HIV. We first predicted this secondary structure based on measurements of selection pressure against synonymous mutations and synonymous mutation covariance. We observed that specific synonymous mutations occur far less frequently than would be expected under a neutral model, indicating that individual viruses containing these mutations reproduced much less successfully than viruses with comparable mutations at other sites. If our selection pressure measurements had a purely spurious basis, predictions based upon them should also be spurious. The fact that their predicted base-paired regions have been validated by the SHAPE experiment suggests that they represent real selection pressures. All of the available evidence indicates that even a single mutation disrupting this base-pairing will have much lower reproductive success (as measured by our data from wild HIV populations) than comparable synonymous mutations at other nearby sites. This is evidence of an important function, but with no suggestion of what that function might be.

There are a variety of computational methods (Hofacker et al. 2002; Knudsen and Hein 2003; Pedersen et al. 2004) that combine covariant substitution models with the thermodynamic prediction methods. These methods are superior if we have only a small set of sequences, where the evidence from covariant substitutions can hardly infer the secondary structure by itself. In our analysis, the large sequence data set enables the covariance information to serve as independent evidence for the predicted structure, predicting individual base pairs with significant P-values.

Experimental detection of RNA secondary structure using SHAPE

SHAPE appears to be an effective and efficient new approach for detecting RNA secondary structure. Based on the principle that unpaired regions are more accessible to NMIA acylation (Merino et al. 2005), it exploits the fact that acylated nucleotides will terminate reverse transcription to yield a result that can be easily read out like a sequencing gel. Originally developed by Weeks and coworkers, it has been applied to a variety of problems including studies of tRNA and the dimerization domain of a retroviral genome (Badorrek and Weeks 2005; Wilkinson et al. 2005), yielding results consistent with established methods of structure analysis. Another study has demonstrated the accuracy of a SHAPE analysis of the telomerase RNA using NMR (Chen et al. 2006). Recently, SHAPE was applied to a different section of the HIV-1 genome, the first 906 nt of the 5′ noncoding region (Wilkinson et al. 2006), revealing several secondary structural elements. These studies indicate that SHAPE is a robust and widely applicable technique.

There is a striking level of agreement between the SHAPE results, which measure nucleotide accessibility and conformation versus the synonymous mutation and covariance results, which measure evolutionary selection pressure. They both give strong evidence supporting stems A and B (Figs. 3–5). In contrast, at stem C, the most flexible stem structure according to SHAPE, there is no evidence of conservation in comparative sequence analyses. First, stem C shows no significant reduction in synonymous variability vs. neighboring regions (P-value = 0.9). Second, no significant mutation covariation was detected in stem C (Fig. 4). Third, among the 37 reference sequences of HIV-1 group M, the stem C is the least conserved stem in the predicted structure (Supplemental Table 1).

Role of the RNA secondary structure

What is the possible biological function of this secondary structure? This is largely unknown. One possibility in the literature is that it might assist recombination. The predicted structure is similar to a secondary structure whose role in recombination in HIV has been well studied (Galetto et al. 2004). They both contain an ∼20-bp-long stem at the base (Fig. 2, stem A versus S1 by Galetto et al. [2004]); they both contain a 7-9 bp stem (Fig. 2, stem B versus S2 by Galetto et al. [2004]) less than 10 nt downstream of stem A or S1, with a bubble on the end of the stem B or S2. In addition, the well-studied structure contains an 18-nt bubble (Galetto et al. 2004, L1) downstream of stem S2, while downstream of the corresponding stem B there is also a bubble of 16 nt at the end of stem C. In contrast to the similarity between our predicted structure and the structure known to assist recombination, the predicted structure is less similar to the secondary structures in HIV with other functions, such as TAR (Berkhout 1992) and the *gag*/*pol* frame-shift hairpin (Parkin et al. 1992), both of which are single stem–loop structures, as well as RRE (Malim et al. 1989), which is a more complicated structure with compounded stem–loop structures.

Based on the example of the S1/S2 structure (Galetto et al. 2004), one might predict that a recombination hot spot could be found within our proposed structure (Galetto et al. 2006). Recombination breakpoints have been found in nucleotide region 230–330, i.e., in the sequence that folds to the predicted structure, between the *pol* gene sequences from different subtypes, such as between subtypes A and B, A and D, and B and D, as well as between B and F (Quarleri et al. 2004; Yang et al. 2004; Sa Filho et al. 2005). However, the existence of breakpoints in this region is insufficient to define a recombination hot spot. Laboratory experiments need to be designed specifically to test the hypothesis that the recombination frequency in this region is higher than that of the neighboring regions (Moumen et al. 2001).

However, even if this explanation were correct, it seems unlikely that recombination alone could explain the very strong levels of selection pressure observed in this sequence region, implying the possibility of other, more important functions. An obvious first step would be to measure the effect of these synonymous mutations on HIV reproductive success in standard laboratory assays. If this secondary structure is shown to be important for the viral life cycle in the laboratory, it will be both interesting and possible to dissect its functional importance experimentally.

MATERIALS AND METHODS

HIV-1 subtype B sequence data

This data set contained just subtype B sequences, mostly from patients under antiretroviral drug treatment (Chen et al. 2004; Pan et al. 2007). These sequences cover 450 codons, including the whole protease (99 codons) and the first 351 codons of the reverse transcriptase. Multiple sequence alignments and mutation detection were performed as previously described (Chen et al. 2004). To ensure sufficient diversity of the data and minimize potential bias from phylogenetic effects, we excluded sequences with less than 2% nucleotide diversity. After filtering, there are 20,042 sequences available, which have been included in this analysis.

HIV-1 consensus sequences and reference sequences

The consensus sequence of HIV-1 group M was downloaded from the Los Alamos HIV Sequence Database version August 2004. Aligned HIV-1 group M reference sequences were downloaded from HIV-1 Subtype Reference Alignments in the version 2007 of Los Alamos HIV Sequence Database.

Thermodynamic prediction, free energy calculation, and base-pair probabilities

RNAfold in the Vienna package version 1.6.1 and mfold version 3.2 (http://frontend.bioinfo.rpi.edu/) were used to predict structures using the default parameters. RNAfold was also used to measure the MFE for each sequence with its default parameters. It predicts the free energy of the most stable RNA structure for a given sequence. The base-pair probabilities were calculated by RNAfold as well (McCaskill 1990).

Sequence shuffling tests

All sequence randomization was carried out using the Dicodon-Shuffle algorithm (Katz and Burge 2003), which retains the dinucleotide composition at the (3,1), (1,2), and (2,3) positions as well as the encoded amino acid sequence and codon usage of the native coding sequence. The program was downloaded from the authors' website at ftp://hollywood.mit.edu.

Measure of synonymous variability

At each amino acid position, we calculated the normalized mean pairwise distance (MPD) as described (Gog et al. 2007). The MPD, as a measure of sequence variability, is the sum of all individual pairwise distances divided by the number of pairs [n sequences give $n(n-1)/2$ possible sequence pairs]. At each amino acid position, we only analyzed codons coding the most frequent amino acid, so that the difference between any two codons is synonymous. Hence, these MPD scores do not reflect the amino acid-level selection (Simmonds and Smith 1999). To take codon bias into consideration, the MPD score is normalized by the expected MPD given the distribution of codons for that amino acid in the whole segment (Gog et al. 2007). At positions with invariant tryptophan and methionine codons, the expected MPDs are zero. Positions with these amino acids are not included in our moving average calculation.

Measure of covariance

We used Fisher's exact test (Fisher 1922; Agresti 1992) to test for nonrandom associations between mutation α at position X and mutation β at position Y, by computing the P-value for the two-sided test using the 2×2 contingency table: $N_{X\alpha Y\beta}$, $N_{X\alpha Y0}$, $N_{X0Y\beta}$, and N_{X0Y0}. $N_{X\alpha Y\beta}$ is the number of samples that have mutation α at position X and also mutation β at position Y; $N_{X\alpha Y0}$ is the number of samples that have mutation α at position X but no mutation at position Y; $N_{X0Y\beta}$ is the number of samples that have no mutation at position X and have mutation β at position Y; N_{X0Y0} is the number of samples that have mutations at neither position. We computed the odds ratio, its confidence interval (95% two-sided), and the P-value using the fisher.test function from the statistical software package R. The lower-bound estimate for the strength of covariation, based on a 95% confidence interval, is defined as the covariance score.

Covariation map

The covariation map was derived as described (Wang and Lee 2007). Only statistically significant mutation pairs ($P < 10^{-8}$ for a single pair by Fisher's exact test, yielding a significance level of 0.003 for each pair after the Bonferroni correction for all possible synonymous mutation pairs) were included in our analysis. For a given codon position pair, the strongest covariance score for any pair of synonymous mutations at the two positions was displayed in the map.

Chi-square test and Fisher's exact test for covariant sites involving multiple pairs of mutations

The chi-square test (the chisq.test function from the statistical software package R) was used to test for nonrandom associations between synonymous mutations at position *X* with those at position *Y*. The contingency table used in the chi-square test has two variables for positions *X* and *Y*, respectively, each including all possible synonymous mutations at that position. The number in the contingency table where *X* is α and *Y* is β indicates the number of samples having synonymous difference α at position *X* and synonymous difference β at position *Y*.

Fisher's exact test (the fisher.test function from the statistical software package R) was used on the same contingency table, except the columns and rows associated with the wild type (consensus), because the program cannot handle the large numbers in these columns and rows.

Hypergeometric significance test for enrichment of covariant mutation pairs

Assuming there are *M* pairs of strongly covariant mutations out of a total of *N* pairs and we observed *m* pairs of strongly covariant mutations in a subset containing *n* pairs, the statistical significance that the subset is enriched with strongly covarying pairs is inferred by the probability that *m* or more strongly covarying pairs are observed in the subset under a random model. Under the random model, if we randomly draw *n* pairs from the total set of *N* pairs, the probability that *m* of them are strongly covarying pairs follows the Hypergeometic distribution (dhyper function in the statistical software package R).

SHAPE analysis

Cloning and transcription

The SHAPE method (Merino et al. 2005; Wilkinson et al. 2006) was used to probe unpaired regions in the predicted secondary structure. A 113-nt sequence fragment, nucleotides 1833–1945 of a reference genome of HIV-1 subtype A1 (GenBank accession number: U51190), was synthesized using in vitro transcription. The RNA was cloned into the pUC19 vector between EcoRI and XbaI restriction sites. The construct contains, in addition to the HIV sequence, a 5′ linker and a 3′ linker containing the RT primer binding site (Wilkinson et al. 2006). This cloned plasmid was linearized with XbaI and used in run-off transcription by T7 RNA Polymerase (Milligan and Uhlenbeck 1989). Transcriptions were precipitated in ethanol and purified on a 6% denaturing polyacrylamide gel (29:1 acrylamide:bis-acrylamide). The RNA was resuspended in 0.5× TE buffer (5 mM Tris at pH 7.5, 0.5 mM EDTA).

Radiolabeling

Reverse transcription primer from IDT (Integrated DNA Technologies) was labeled using γ^{32}P ATP and polynucleotide kinase. The labeled, full-length primer was purified on a 20% denaturing polyacrylamide gel. The primer was ethanol precipitated before use in reverse transcription.

NMIA modification

We followed the method of Wilkinson et al. (2006) with a few modifications. We annealed 12 pmol of RNA at 95°C for 3 min, then chilled it at 4°C for 1 min. We then added 3.33× annealing buffer (333 mM HEPES at pH 8.0, 20 mM $MgCl_2$, 333 mM NaCl) and incubated at 37°C for 20 min. The annealed RNA was treated with 6 mM NMIA in DMSO or DMSO alone for 45 min at 37°C, then ethanol precipitated and reverse transcribed.

Sequencing gel running and analysis

Sequencing reactions were run on a 12% denaturing polyacrylamide gel. The sequencing gels were run at 60 W for the indicated times and dried at 80°C for 60 min, and then exposed to a storage phosphor screen for 72 h. The gel images were produced using a Typhoon phosphorimager (GE Healthcare). Analysis and quantification were carried out using the SAFA program (Das et al. 2005). Band intensities in the NMIA (−) lane were subtracted from the (+) lane in order to determine the contribution of NMIA to the signal in the (+) lane, and the highest degree of NMIA modification was set to 100%.

SUPPLEMENTAL DATA

Supplemental material can be found at http://www.rnajournal.org.

ACKNOWLEDGMENTS

We thank Alex Alekseyenko, Mijeong Kang, Aron Yoffe, and the three anonymous reviewers for helpful comments and discussions on the manuscript; Art Poon for helping to calculate the covariation controlling for phylogeny; and Frederick Bibollet-Ruche, Matteo Negroni, Frank Maldarelli, and Paul Sharp for helpful discussions on the potential function of the structure. This work was supported by grants from the NIH (U54 RR021813) and DOE (DE-FC02-02ER63421), a Dreyfus Foundation Teacher-Scholar Award to C. L., and UCLA dissertation year fellowship and UCLA AIDS Institute fellowship (UCLA AIDS Institute, and the UCLA Center for AIDS Research [AI28697]) to Q.W.

Received January 24, 2008; accepted September 24, 2008.

REFERENCES

Abbink, T.E. and Berkhout, B. 2003. A novel long distance base-pairing interaction in human immunodeficiency virus type 1 RNA occludes the Gag start codon. *J. Biol. Chem.* **278:** 11601–11611.

Agresti, A. 1992. A survey of exact inference for contingency tables. *Stat. Sci.* **7:** 131–177.

Badorrek, C.S. and Weeks, K.M. 2005. RNA flexibility in the dimerization domain of a γ retrovirus. *Nat. Chem. Biol.* **1:** 104–111.

Berkhout, B. 1992. Structural features in TAR RNA of human and simian immunodeficiency viruses: A phylogenetic analysis. *Nucleic Acids Res.* **20:** 27–31.

Chen, Y., Carlini, D.B., Baines, J.F., Parsch, J., Braverman, J.M., Tanda, S., and Stephan, W. 1999. RNA secondary structure and compensatory evolution. *Genes Genet. Syst.* **74:** 271–286.

Chen, L., Perlina, A., and Lee, C.J. 2004. Positive selection detection in 40,000 human immunodeficiency virus (HIV) type 1 sequences automatically identifies drug resistance and positive fitness mutations in HIV protease and reverse transcriptase. *J. Virol.* **78:** 3722–3732.

Chen, Y., Fender, J., Legassie, J.D., Jarstfer, M.B., Bryan, T.M., and Varani, G. 2006. Structure of stem–loop IV of *Tetrahymena* telomerase RNA. *EMBO J.* **25:** 3156–3166.

Damgaard, C.K., Andersen, E.S., Knudsen, B., Gorodkin, J., and Kjems, J. 2004. RNA interactions in the 5′ region of the HIV-1 genome. *J. Mol. Biol.* **336:** 369–379.

Das, R., Laederach, A., Pearlman, S.M., Herschlag, D., and Altman, R.B. 2005. SAFA: Semiautomated footprinting analysis software for high-throughput quantification of nucleic acid footprinting experiments. *RNA* **11:** 344–354.

Fisher, R.A. 1922. On the interpretation of $\chi(2)$ from contingency tables, and the calculation of *P. J. R. Stat. Soc. [Ser A]* **85:** 87–94.

Galetto, R., Moumen, A., Giacomoni, V., Veron, M., Charneau, P., and Negroni, M. 2004. The structure of HIV-1 genomic RNA in the gp120 gene determines a recombination hot spot in vivo. *J. Biol. Chem.* **279:** 36625–36632.

Galetto, R., Giacomoni, V., Veron, M., and Negroni, M. 2006. Dissection of a circumscribed recombination hot spot in HIV-1 after a single infectious cycle. *J. Biol. Chem.* **281:** 2711–2720.

Gog, J.R., Afonso Edos, S., Dalton, R.M., Leclercq, I., Tiley, L., Elton, D., von Kirchbach, J.C., Naffakh, N., Escriou, N., and Digard, P. 2007. Codon conservation in the influenza A virus genome defines RNA packaging signals. *Nucleic Acids Res.* **35:** 1897–1907.

Gruber, A.R., Neubock, R., Hofacker, I.L., and Washietl, S. 2007. The RNAz web server: Prediction of thermodynamically stable and evolutionarily conserved RNA structures. *Nucleic Acids Res.* **35:** W335–W338.

Hofacker, I.L., Fekete, M., Flamm, C., Huynen, M.A., Rauscher, S., Stolorz, P.E., and Stadler, P.F. 1998. Automatic detection of conserved RNA structure elements in complete RNA virus genomes. *Nucleic Acids Res.* **26:** 3825–3836.

Hofacker, I.L., Fekete, M., and Stadler, P.F. 2002. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.* **319:** 1059–1066.

Hofacker, I.L., Stadler, P.F., and Stocsits, R.R. 2004. Conserved RNA secondary structures in viral genomes: A survey. *Bioinformatics* **20:** 1495–1499.

Katz, L. and Burge, C.B. 2003. Widespread selection for local RNA secondary structure in coding regions of bacterial genes. *Genome Res.* **13:** 2042–2051.

Knudsen, B. and Hein, J. 2003. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res.* **31:** 3423–3428.

Malim, M.H., Hauber, J., Le, S.Y., Maizel, J.V., and Cullen, B.R. 1989. The HIV-1 rev *trans*-activator acts through a structured target sequence to activate nuclear export of unspliced viral mRNA. *Nature* **338:** 254–257.

Mathews, D.H. and Turner, D.H. 2006. Prediction of RNA secondary structure by free energy minimization. *Curr. Opin. Struct. Biol.* **16:** 270–278.

McCaskill, J.S. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **29:** 1105–1119.

Merino, E.J., Wilkinson, K.A., Coughlan, J.L., and Weeks, K.M. 2005. RNA structure analysis at single nucleotide resolution by selective 2′-hydroxyl acylation and primer extension (SHAPE). *J. Am. Chem. Soc.* **127:** 4223–4231.

Milligan, J.F. and Uhlenbeck, O.C. 1989. Synthesis of small RNAs using T7 RNA polymerase. *Methods Enzymol.* **180:** 51–62.

Moumen, A., Polomack, L., Roques, B., Buc, H., and Negroni, M. 2001. The HIV-1 repeated sequence R as a robust hot-spot for copy-choice recombination. *Nucleic Acids Res.* **29:** 3814–3821.

Ooms, M., Abbink, T.E., Pham, C., and Berkhout, B. 2007. Circularization of the HIV-1 RNA genome. *Nucleic Acids Res.* **35:** 5253–5261.

Paillart, J.C., Skripkin, E., Ehresmann, B., Ehresmann, C., and Marquet, R. 2002. In vitro evidence for a long range pseudoknot in the 5′-untranslated and matrix coding regions of HIV-1 genomic RNA. *J. Biol. Chem.* **277:** 5995–6004.

Pan, C., Kim, J., Chen, L., Wang, Q., and Lee, C. 2007. The HIV positive selection mutation database. *Nucleic Acids Res.* **35:** D371–D375.

Parkin, N.T., Chamorro, M., and Varmus, H.E. 1992. Human immunodeficiency virus type 1 *gag-pol* frameshifting is dependent on downstream mRNA secondary structure: Demonstration by expression in vivo. *J. Virol.* **66:** 5147–5151.

Pedersen, J.S., Meyer, I.M., Forsberg, R., Simmonds, P., and Hein, J. 2004. A comparative method for finding and folding RNA secondary structures within protein-coding regions. *Nucleic Acids Res.* **32:** 4925–4936.

Peleg, O., Brunak, S., Trifonov, E.N., Nevo, E., and Bolshoy, A. 2002. RNA secondary structure and squence conservation in C1 region of human immunodeficiency virus type 1 env gene. *AIDS Res. Hum. Retroviruses* **18:** 867–878.

Peleg, O., Trifonov, E.N., and Bolshoy, A. 2003. Hidden messages in the nef gene of human immunodeficiency virus type 1 suggest a novel RNA secondary structure. *Nucleic Acids Res.* **31:** 4192–4200.

Quarleri, J.F., Rubio, A., Carobene, M., Turk, G., Vignoles, M., Harrigan, R.P., Montaner, J.S., Salomon, H., and Gomez-Carrillo, M. 2004. HIV type 1 BF recombinant strains exhibit different pol gene mosaic patterns: Descriptive analysis from 284 patients under treatment failure. *AIDS Res. Hum. Retroviruses* **20:** 1100–1107.

Sa Filho, D.J., Sanabani, S., Diaz, R.S., Munerato, P., Brunstein, A., Fusuma, E., Sabino, E.C., and Janini, L.M. 2005. Analysis of full-length human immunodeficiency virus type 1 genome reveals a variable spectrum of subtypes B and f recombinants in Sao Paulo, Brazil. *AIDS Res. Hum. Retroviruses* **21:** 145–151.

Schinazi, R.F., Lloyd Jr., R.M., Ramanathan, C.S., and Taylor, E.W. 1994. Antiviral drug resistance mutations in human immunodeficiency virus type 1 reverse transcriptase occur in specific RNA structural regions. *Antimicrob. Agents Chemother.* **38:** 268–274.

Simmonds, P. and Smith, D.B. 1999. Structural constraints on RNA virus evolution. *J. Virol.* **73:** 5787–5794.

Simmonds, P., Tuplin, A., and Evans, D.J. 2004. Detection of genome-scale ordered RNA structure (GORS) in genomes of positive-stranded RNA viruses: Implications for virus evolution and host persistence. *RNA* **10:** 1337–1351.

Simon, V., Ho, D.D., and Abdool Karim, Q. 2006. HIV/AIDS epidemiology, pathogenesis, prevention, and treatment. *Lancet* **368:** 489–504.

van Hemert, F.J. and Berkhout, B. 1995. The tendency of lentiviral open reading frames to become A-rich: Constraints imposed by viral genome organization and cellular tRNA availability. *J. Mol. Evol.* **41:** 132–140.

Wang, Q. and Lee, C. 2007. Distinguishing functional amino acid covariation from background linkage disequilibrium in HIV protease and reverse transcriptase. *PLoS One* **2:** e814. doi: 10.1371/journal.pone.0000814.

Wilkinson, K.A., Merino, E.J., and Weeks, K.M. 2005. RNA SHAPE chemistry reveals nonhierarchical interactions dominate equilibrium structural transitions in tRNA(Asp) transcripts. *J. Am. Chem. Soc.* **127:** 4659–4667.

Wilkinson, K.A., Merino, E.J., and Weeks, K.M. 2006. Selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE): Quantitative RNA structure analysis at single nucleotide resolution. *Nat. Protocols* **1:** 1610–1616.

Xing, Y. and Lee, C. 2006. Can RNA selection pressure distort the measurement of Ka/Ks? *Gene* **370:** 1–5.

Yang, C., Li, M., Shi, Y.P., Winter, J., van Eijk, A.M., Ayisi, J., Hu, D.J., Steketee, R., Nahlen, B.L., and Lal, R.B. 2004. Genetic diversity and high proportion of intersubtype recombinants among HIV type 1-infected pregnant women in Kisumu, western Kenya. *AIDS Res. Hum. Retroviruses* **20:** 565–574.

Zuker, M. 1989. On finding all suboptimal foldings of an RNA molecule. *Science* **244:** 48–52.

APPENDIX A

Appendices and Data Tables

A.1 Measuring degree of cooperativity using the Hill equation and Hill Plot

A.1.1 Derivation of the Hill equation.

The Hill equation [70, 71, 72] gives us a method of estimating the lower limit of the number of subunits that associate during a reaction, the *order* of binding. The model is used when the first binding event increases the probability of a second binding event, and so on, making the binding *cooperative*. Mechanistically, the assumption is that the association is one-step, with n protein units associating at the same time. This approximates a situation where the first subunit binding is the slow step, and the subsequent binding events occur rapidly, so it can be though of as a single step - even if it isn't, mechanistically speaking. This assumption is important to keep in mind when interpreting the plot.

Consider a case where we have n identical proteins binding to a strand of RNA. The fraction of RNA (R) bound to Protein (P) is given by

$$F = \frac{P_nR}{R + P_nR}$$

which rearranges to

$$\frac{R}{P_nR} = \frac{1}{F} - 1 \tag{A.1}$$

The dissociation constant for the reaction $nP + R \rightleftharpoons P_nR$ is

$$K_D = \frac{P^nR}{P_nR} \,.$$

This dissociation constant is equal to the monomolecular dissociation constant raised to the n-th power, so $K_D = K^n$. Substituting Equation A.1 from above, it rearranges to

$$F = \frac{P^n}{K^n + P^n} \tag{A.2}$$

It is also important to note that this model assumes that the amount of unbound protein P is roughly equal to the total amount of protein in the reaction, which will only be true in the case where $R \ll P$. In the Hill plot, the y axis is defined as the logit function of F,

$$y = \text{logit}(F) = \log\left(\frac{F}{1-F}\right)$$

so, substituting in Equation A.2 gives

$$y = \log\frac{P^n}{K^n}$$

Since the x axis is $\log P$,

$$y = n\log P - n\log K = nx - n\log K \tag{A.3}$$

which is the equation for a straight line with the Hill constant as the slope, and the x intercept is

$$x_{int} = \log K \tag{A.4}$$

The Hill constant and the K of the complex can be estimated by a linear regression of the data in the transition region of the Hill plot, near where $P \approx K$.

A.1.2 Uses of the Hill plot

The Hill plot is used to estimate the cooperativity of a binding reaction. Ideally, it would give a straight line for a system that worked exactly according to the assumptions: after one subunit binds, it would immediately recruit the other $n-1$ subunits, which would bind more or less instantaneously. In reality, if the higher order binding steps have an affinity close to that of the first, the slope will deviate from a straight line on the Hill plot and the measured Hill constant will be reduced. Because of this, the line will only approach the correct order of binding in the limit; for example, an experimentally determined Hill constant of 2.6 suggests an order of binding of 3 or higher. One thing to remember is that the model shown above conflicts in a very important way with the assumptions of the Hill plot. The number of subunits, n, is an integer, whereas the slope of the Hill plot determined by linear regression is not [71, 72, 65]. This has lead some authors (e.g. [65]) to conclude that the derivation shown above, using the equation $nP + R \rightleftharpoons P_nR$, is incorrect. However, interpretation of n as a degree of cooperativity instead of an integer, or as a poor approximation of an integer, saves this interpretation to a small degree. The derivation given also has the pedagogical advantage of relating the degree of cooperativity to the slope of the Hill plot, a connection that is less clear if we treat Equation A.2 as simply an empirical relationship with no physical significance.

One other important consideration when using the Hill equation is that the protein concentrations used must be highly accurate, or the final data will be poor. Differentiating Equation A.2 with respect to protein concentration gives

$$\left(\frac{\partial F}{\partial P}\right)_n = \frac{K^n n P^{n-1}}{(K^n + P^n)^2}.$$

And the second derivative is

$$\left(\frac{\partial^2 F}{\partial P^2}\right)_n = \frac{K^n P^{n-1}(K^n + P^n - (P^n - K^n) n \log P)}{(K^n + P^n)^3}.$$

This function has a zero where $P = K\left(\frac{n-1}{n+1}\right)^{1/n}$, indicating that the effect of pipetting error on the individual data points is greatest at around the dissociation constant. When looking at the data, this appears as a series of well-behaved, consistent points where F is close to 0 or 1, but strong fluctuations where $P \sim K$. This is unfortunate, given that the data points around K are the most essential ones for determining the correct n and K; the only remedy is to make the dilution series carefully.

A.2 Pyridine Hemochromagen assay

A.2.1 Historical Considerations

G.G. Stokes was the first to prepare what we now refer to as hemochromagen. As early as 1863, he was monitoring changes of the hemoglobin absorbance spectrum upon reduction of the heme to the Fe(II) form. Stokes had reduced blood in the presence of ammonia; what he was seeing were the intense α and β peaks of Fe(II) heme b complexed with ammonia. Later authors [73] were able to show that the hemochromagen, as it had been called by Christian Bohr [74], was heme in complex with some nitrogenous ligand. It can be formed as well by simply reducing myoglobin under denaturing conditions, in which case a histidine serves as the axial ligand.

With advances in spectroscopic techniques, later workers were able to use hemochromagen for an important analytical purpose [75]. The regularity of the α peak of reduced heme and its high extinction coefficient allowed its use for determining the total heme composition in a sample. De Duve [76] published one of the first protocols for this, and determined a value for the absolute absorbance at 557 nm. Paul, Theorell, and Åkeson [77], after noticing that the pyridine assay done with values similar to De Duve's gave inconsistent results, redetermined the extinction coefficient for for pyridine hemochromes from recrystallized heme, non-recrystallized heme, and myoglobin, using the iron content as an internal control. They found a value (34.7 $\text{cm}^{-1}\text{mM}^{-1}$) significantly higher than the previous value of De Duve (32 $\text{cm}^{-1}\text{mM}^{-1}$), representing a difference of roughly 9%. However, both values are unfortunately still in use today (e.g. [78], which cites Paul et al.'s coefficient for the absolute spectrum, but uses a value for the difference spectrum that was derived from De Duve's value). Given that the combined error of dilution and spectrophotometric measurement is easily below 5% using modern equipment, it is a shame that such confusion exists.

A.2.2 Pyridine Hemochromagen assay: Experimental Method

Based on the paper by Berry and Trumpower (1987) *Analytical Biochemistry*, **161**, 1-15.

The volumes shown below can be scaled up or down as desired, as long as the concentrations stay the same. The concentration of heme in your sample should be at least 10 μM and not greater than 80 μM.

We make a stock containing 0.2 M NaOH, 40% (v/v) Pyridine; we add 0.5 mL of this to a tube, plus 3 μL of 0.1 M Potassium Ferricyanide ($K_3Fe(III)(CN)_6$). To this we add 0.5 mL of our sample, mix well, and scan as soon as possible. Pyride has an unpleasant smell, so it's best to use a sealed cuvette and work in the fume hood. We will need to scan the region between 500 and 600 nm; also, set the spectrophotometer to $\leq$ 0.5 nm bandwidth and collect data at $\leq$ 1 nm interval. Record the first scan, the **oxidized scan**. What has happened is that the Potassium Ferricyanide has ensured that all the heme in solution is in the Fe(III) form. We will now reduce the heme and scan again.

Add a few crystals of Sodium Dithionite to the sample and mix thoroughly. You might see the sample change color as the intense bands in the visible region appear upon reduction. The Fe(III) in the sample is now reduced to Fe(II), and the heme has taken up pyridine as a ligand. Scan immediately; this is the first **reduced scan**. Then, wait 1 min and scan again. You may see the peak increase. Keep scanning at minute intervals and the peak will eventually begin to decrease. Take the maximum peak as your reduced scan.

The concentration of heme in the sample can be determined from the absorbance of the reduced sample; the extinction coefficient (ε) for reduced pyridine hemochromagen is 34.53 mM $^{-1}$cm^{-1} at 557 nm. It is also possible to subtract the oxidized spectrum from the reduced spectrum to control for turbidity of the sample. First, subtract the absorbances of the oxidized spectrum from the corresponding reduced spectrum values. Then, as recommended in the paper, subtract the value at 540 nm (which should be at the low point between the peaks at 525 and 557 nm) from the value at 557 nm. The extinction coefficient for this value is 23.98 mM $^{-1}$cm^{-1}. Don't forget, in both cases, to take your dilution factor into consideration.

When you know your heme concentration, take a good scan of your native protein. Since you now know the concentration of heme, divide the absorbances of the heme spectrum by the heme concentration to get the extinction coefficients for your hemoprotein at wherever you wish, since $\varepsilon = A/(c \cdot l)$. This is not valid in regions where other chromophores absorb, e.g. at < 310 nm.

Hazards: Pyridine is toxic, flammable, and can cause sterility in males. Mix solutions in a fume hood. Sodium dithionite is flammable. Potassium Ferricyanide is an oxidant. **Warnings:** Pyridine eventually goes bad when contaminated with water. When this happens it turns yellow. Do not use if yellow, and seal the pyridine bottle with Parafilm after use. Sodium dithionite should also be kept from water. Potassium Ferricyanide should be made fresh from dry powder and used the same day. You only need a small amount, so make a stock of less than 10 mL. It should be disposed of in its own waste container.

A.2.3 Error Analysis of the Method

The extinction coefficient of heme in a hemoprotein can be found using the pyridine hemochromagen assay [76, 77, 79]. The formation of highly-absorbing Fe(II) species adduct with pyridine gives an easy way to determine the amount of heme in a sample, and along with the electronic absorbance spectrum of the native protein can give an accurate heme extinction coefficient. The Beer-Lambert law, which relates concentration and measured absorbance, is as follows: $A = \varepsilon \cdot c \cdot l$, i.e. absorbance is concentration (c) times path length (l) times a wavelength dependent factor, the extinction coefficient (ε).

We are also interested in the uncertainty of our measurement, which we will express using the operator δ. For example, δX is the uncertainty in the measurement of X, and $\overline{X}$ is the average value of X, determined experimentally. ε_λ is the extinction coefficient of the pyridine hemochrome at some reference wavelength, A_λ is the absorbance of hemochrome at the reference wavelength. c_H = concentration of heme in the sample, $a(\lambda)$ is the absorbance of native protein at wavelength λ. $\varepsilon(\lambda)$ is the native protein extinction coefficient at wavelength λ. We will assume a constant path length of 1 cm in all cases, with no uncertainty. Given that ε_λ is a fixed parameter, the fractional uncertainty in the heme concentration over the pyridine experiment is

$$\frac{\delta c_H}{\overline{c_H}} = \frac{\delta A_\lambda}{\overline{A_\lambda}}$$

The extinction coefficients of the native protein are determined from the native protein absorbance and the experimentally determined heme concentration,

$$\varepsilon(\lambda) = \frac{a(\lambda)}{c_H}$$

so the fractional uncertainty in the extinction coefficient is

$$\frac{\delta \varepsilon(\lambda)}{\overline{\varepsilon(\lambda)}} = \sqrt{\left(\frac{\delta c_H}{\overline{c_H}}\right)^2 + \left(\frac{\delta a(\lambda)}{\overline{a(\lambda)}}\right)^2}$$

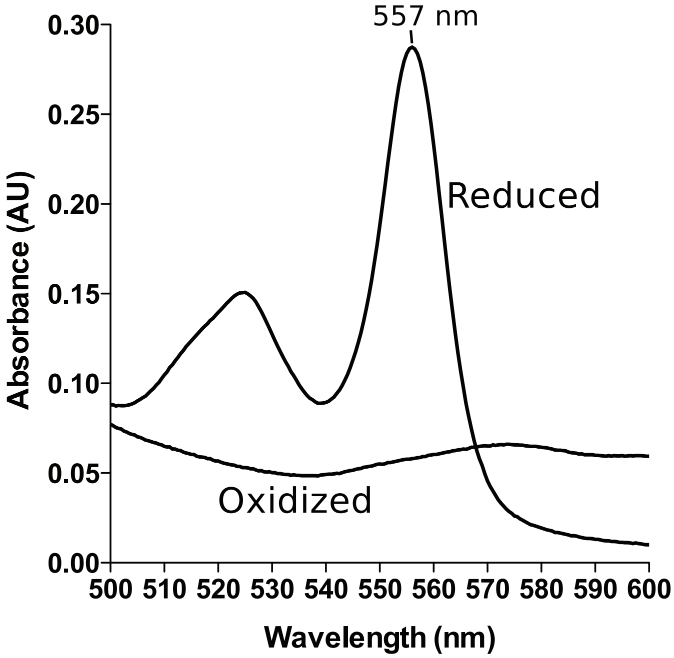

Figure A.1: **Reduced and oxidized Pyridine Hemochromagen.** The example shown is Frog DGCR8 heme-binding domain, scanned on a Cary 300 Biospectrophotometer set at 0.5 nm bandwidth.

Substituting from above and rearranging gives us

$$\delta\varepsilon(\lambda) = \overline{\varepsilon(\lambda)}\sqrt{\left(\frac{\delta A_\lambda}{\overline{A_\lambda}}\right)^2 + \left(\frac{\delta a(\lambda)}{\overline{a(\lambda)}}\right)^2}$$

More specifically, the standard deviation of the extinction coefficient is given by

$$\sigma_{\varepsilon(\lambda)} = \overline{\varepsilon(\lambda)}\sqrt{\left(\frac{\sigma_{A_\lambda}}{\overline{A_\lambda}}\right)^2 + \left(\frac{\sigma_{a(\lambda)}}{\overline{a(\lambda)}}\right)^2}$$

Thus, the uncertainty in the extinction coefficient of the heme at a given wavelength is the calculated value times the fractional uncertainties in both absorbencies, added in quadrature [80, 81, 82].

A.3 An Interpretation of Vierordt's Method For Determining Concentrations of Multiple Substances in Solution

Vierordt's method is a technique of analytical chemistry to determine the concentrations of two or more chromophores in solution by determining their individual extinction coefficients at multiple wavelength and setting up a set of simultaneous equations to solve the unknown concentrations [83]. Vierordt ([32], pg 52) gave the following equations for the concentrations of two substances x and y[1]:

$$x = \frac{(E'd - Eb)ac}{ad - bc},\; y = \frac{(Ea - A'c)bd}{ad - bc} \tag{A.5}$$

He was using another form of Beer's law where the absorbance, denoted by E, equals $E = \frac{x}{a}$. a is therefore closer to the reciprocal of the extinction coefficient. He gives no derivation of Equation A.5, though it is easy to see that it is simply a case of solving a system of two simultaneous equations.

Here I give a method of deriving the concentrations based on a vector interpretation, the main benefit being that it allows one to better visualize the method and see how it works.

Consider an n-dimensional vector space, where each dimension is the extinction coefficient at some wavelength. So, for species i the extinction coefficient vector is

$$\mathbf{E}_i = \begin{bmatrix} \varepsilon_{i,1} \\ \vdots \\ \varepsilon_{i,n} \end{bmatrix}$$

where each $\varepsilon_{i,j}$ is the extinction coefficient of species i at wavelength j. The Absorbance vector $\mathbf{A}_i$ of substance i is given by the concentration c_i, a scalar, multiplied by $\mathbf{E}_i$[2].

$$\mathbf{A}_i = c_i \mathbf{E}_i$$

---

[1] There is a typo in one of his equations; he has $E' = \frac{x}{c} + \frac{y}{\mathbf{a}}$ instead of $E' = \frac{x}{c} + \frac{y}{\mathbf{d}}$

[2] This is a version of the Beer-Lambert law, $A = \varepsilon c l$. Note that I am neglecting the path length l, on the assumption that it will remain constant for all species and all wavelengths. We can consider it as part of the extinction coefficient.

The total measured absorbance vector of the sample at the chosen wavelengths is $\mathfrak{A}$, which is a linear combination of the m different species in the sample.

$$\mathfrak{A} = \sum_{i=1}^{m} \mathbf{A}_i = \sum_{i=1}^{m} c_i \mathbf{E}_i$$

Note that this vector $\mathfrak{A}$ is, literally, the absorbance readings you get from the spectrophotometer at the different wavelengths you choose to monitor.

Proof that the set of extinction vectors must be linearly independent in order for the concentration of the different species to be determined

For a space of m dimensions, I will show that we can at most estimate the concentration of m different components of the measured absorbance vector $\mathfrak{A}$ using several well-known properties of vector spaces [84]. To estimate correctly, it is necessary that none of the m vectors $\mathbf{E}_i$ can be expressed as a linear combination of the others; the chromophores must be *linearly independent* or we will be able to find several concentrations that are compatible with each measured absorbance vector. The proof of this is as follows: imagine that the measured absorbance vector $\mathfrak{A}$ can be expressed as the sum of m non-zero, linearly-independent extinction vectors $\mathbf{E_1}, \ldots, \mathbf{E_m}$ in two different linear combinations,

$$\mathfrak{A} = k_1\mathbf{E_1} + k_2\mathbf{E_2} + \ldots + k_m\mathbf{E_m}$$

$$\mathfrak{A} = l_1\mathbf{E_1} + l_2\mathbf{E_2} + \ldots + l_m\mathbf{E_m}$$

with k_i and l_i substituting for our concentrations. This is the case if there is more than one possible set of concentrations that gives the same absorbance values at the different wavelengths measured. Of course, this is not what we want; we want to be able to determine the one true set of concentrations uniquely. Let's see what this problem is due to. Setting these equations equal we get:

$$k_1\mathbf{E_1} + k_2\mathbf{E_2} + \ldots + k_m\mathbf{E_m} = l_1\mathbf{E_1} + l_2\mathbf{E_2} + \ldots + l_m\mathbf{E_m}$$

Let us say that, in these two possible sets of concentrations, $k_1 \neq l_1$. Solving for $\mathbf{E_1}$ gives

$$\mathbf{E_1} = \sum_{i=2}^{m} \left(\frac{l_i - k_i}{k_1 - l_1} \right) \mathbf{E_i}$$

If $k_1 \neq l_1$, meaning that there are two concentrations of species 1 compatible with a given absorbance, this means that at least one other k_i is not equal to its corresponding l_i in order to maintain the same absorbance readings. In short, a difference in concentration of one species in solution must be compensated for by concentration from at least one other. The term in parentheses is non-zero for all cases in which this difference in k_i and l_i exists, meaning that $\mathbf{E_1}$ can be expressed as a linear combination of at least one other vector in the set, say $\mathbf{E_2}$.

$$\mathbf{E_1} = \left(\frac{l_2 - k_2}{k_1 - l_1} \right) \mathbf{E_2}$$

This means this set of extinction vectors is linearly *dependent*. From this we conclude that measurement of an Absorbance vector $\mathfrak{A}$ will let us uniquely determine the concentrations $c_1, \ldots, c_m$, so long as the extinction coefficient vectors are all linearly *independent*.

This is only an idealization, however. Uncertainties in the measurement limit the application of this principle. It is also important to be able to know the extinction coefficient of the different chromophores *together in solution*, which often proves difficult. In addition, if the extinction coefficient vectors are too similar, it will be difficult to confidently determine the relative contributions of each component of the solution.

Solving for concentrations of absorbing species as a set of simultaneous equations

The measured absorbance can be expressed as a sum of the extinction coefficient vectors for each of the m chemical species, multiplied by its concentration:

$$\mathfrak{A} = c_1\mathbf{E_1} + c_2\mathbf{E_2} + \ldots + c_m\mathbf{E_m}$$

Expressed in terms of the vectors' components, we have:

$$\begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_m \end{bmatrix} = c_1 \begin{bmatrix} \varepsilon_{1,1} \\ \varepsilon_{1,2} \\ \vdots \\ \varepsilon_{1,m} \end{bmatrix} + c_2 \begin{bmatrix} \varepsilon_{2,1} \\ \varepsilon_{2,2} \\ \vdots \\ \varepsilon_{2,m} \end{bmatrix} + \ldots + c_m \begin{bmatrix} \varepsilon_{m,1} \\ \varepsilon_{m,2} \\ \vdots \\ \varepsilon_{m,m} \end{bmatrix}$$

$$= \begin{bmatrix} \varepsilon_{1,1} & \varepsilon_{2,1} & \cdots & \varepsilon_{m,1} \\ \varepsilon_{1,2} & \varepsilon_{2,2} & \cdots & \varepsilon_{m,2} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{1,m} & \varepsilon_{2,m} & \cdots & \varepsilon_{m,m} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{bmatrix}$$

Or, in compact form, $\mathfrak{A} = \mathbb{E}_{i,j}\,\mathbf{C}_i$. The square matrix $\mathbb{E}$ that contains the extinction coefficients has a few important properties. First, if the extinction vectors of each chromophore are linearly independent, the determinant of the matrix will be non zero; i.e., the matrix is *nonsingular*. In addition, multiplication of the concentration vector by the extinction coefficient matrix gives

$$\begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_m \end{bmatrix} = \begin{bmatrix} c_1\varepsilon_{1,1} + c_2\varepsilon_{2,1} + \ldots + c_m\varepsilon_{m,1} \\ c_1\varepsilon_{1,2} + c_2\varepsilon_{2,2} + \ldots + c_m\varepsilon_{m,2} \\ \vdots \\ c_1\varepsilon_{1,m} + c_2\varepsilon_{2,m} + \ldots + c_m\varepsilon_{m,m} \end{bmatrix}$$

which can be treated as a set of m simultaneous equations of the form $A_\lambda = \sum_i^m c_i \varepsilon_{i,\lambda}$, where λ is a certain wavelength. These simultaneous equations can then be solved using Gaussian elimination, Cramer's rule, or a similar technique.

An Example of the Method

Faller et al. [28] gave the following equation for the number of protein subunits (P) associated with an RNA (R) in a Protein-RNA complex (*complex*), derived using an alternative method:

$$n = \frac{B_R - B_{complex}}{B_R(B_{complex} - B_P)} \frac{\varepsilon_{R,260}}{\varepsilon_{P,280}} \tag{A.6}$$

where $B_i = A_{i,260}/A_{i,280}$. The extinction coefficient matrix of this system is

$$\mathbb{E} = \begin{array}{r} \\ 260 \text{ nm} \\ 280 \text{ nm} \end{array} \begin{array}{c} \begin{array}{cc} RNA & Protein \end{array} \\ \begin{pmatrix} \varepsilon_{R,260} & \varepsilon_{P,260} \\ \varepsilon_{R,280} & \varepsilon_{P,280} \end{pmatrix} \end{array}$$

with the rows and columns labeled for ease of understanding. Using our technique and solving the simultaneous equations

$$\begin{bmatrix} A_{260} \\ A_{280} \end{bmatrix} = \begin{bmatrix} \varepsilon_{R,260} & \varepsilon_{P,260} \\ \varepsilon_{R,280} & \varepsilon_{P,280} \end{bmatrix} \begin{bmatrix} c_R \\ c_P \end{bmatrix}$$

for c_R and c_P using Cramer's rule gives

$$c_R = \frac{\begin{vmatrix} A_{260} & \varepsilon_{P,260} \\ A_{280} & \varepsilon_{P,280} \end{vmatrix}}{\det \mathbb{E}} = \frac{A_{260}\ \varepsilon_{P,280} - A_{280}\ \varepsilon_{P,260}}{\varepsilon_{R,260}\ \varepsilon_{P,280} - \varepsilon_{R,280}\ \varepsilon_{P,260}}$$

$$c_P = \frac{\begin{vmatrix} \varepsilon_{R,260} & A_{260} \\ \varepsilon_{R,280} & A_{280} \end{vmatrix}}{\det \mathbb{E}} = \frac{A_{280}\ \varepsilon_{R,260} - A_{260}\ \varepsilon_{R,280}}{\varepsilon_{R,260}\ \varepsilon_{P,280} - \varepsilon_{R,280}\ \varepsilon_{P,260}}$$

$$n = \quad c_P/c_R \quad = \frac{A_{280}\ \varepsilon_{R,260} - A_{260}\ \varepsilon_{R,280}}{A_{260}\ \varepsilon_{P,280} - A_{280}\ \varepsilon_{P,260}}$$

Given that $B_i = \varepsilon_{i,260}/\varepsilon_{i,280}$ [3], we can derive Equation A.6 above after a little algebra. This method is generalizable to the case of three wavelengths and a 3×3 extinction matrix as well.

If $B_R = B_P$, Equation A.6 is undefined due to the fact that $B_{complex}$ must also equal B_R and B_P, and both the numerator and denominator become zero. If the two chromophores have equal extinction ratios, then any possible mixture of the two will also have the same extinction ratio. In terms of our vector interpretation, if $\varepsilon_{R,260}/\varepsilon_{R,280} = \varepsilon_{P,260}/\varepsilon_{P,280}$, then

$$\begin{bmatrix} \varepsilon_{R,260} \\ \varepsilon_{R,280} \end{bmatrix} = k \begin{bmatrix} \varepsilon_{P,260} \\ \varepsilon_{P,280} \end{bmatrix}$$

where k is some constant, and the extinction vectors are therefore not linearly independent. However, in the case of the protein-RNA complex the ratios are not equal, thus quantification is possible.

Notes for Use

When using in a high-throughput application, say a drug screen, where additional chromophores (or fluorophores) are present, it is important to monitor at least one additional wavelength in order to determine the presence of interfering species. If you have an assay which depends on monitoring wavelengths 1 and 2 to get the relative proportions of species A and B, you should have a fairly robust assay assuming there are no other chromophores present. However, if there is a contaminating species, it is impossible to detect this simply by monitoring two wavelengths. Only by monitoring three can we determine whether or not there is an additional chromophore/fluorophore in the solution. This is especially important in high-throughput screens, where the species are often not well characterized.

---

[3] The paper uses a ratio of absorbances, but for a pure species this is equivalent to a ratio of extinction coefficients, so the distinction is unimportant

Another case in which this technique is useful is when doing analytical chromatography. Often we are limited to detecting at two or three wavelengths simultaneously, so a method like the one shown above is appropriate. For complex reactions in a vessel, however, determination of the entire spectrum of the sample in conjunction with peak-fitting software should give better results. It must be remembered that the technique described above was developed at a time when scanning spectrophotometers were non-existent and wavelengths were difficult to calibrate.

A.4 Data Tables Used in this Thesis

Here I have collected several published extinction coefficients of porphyrin derivatives and molecules of biochemical interest. I have also collected the DGCR8 heme extinction coefficients determined by Rachel Senturia and myself and included in her paper [85]. The table of PPIX derivatives is, to my knowledge, the most complete yet assembled, so I believe it is of some utility to researchers in the field. One special note is that, for certain PPIX derivatives such as Co(III)PPIX-pyridine and Fe(II)PPIX-pyridine, more than one value exists in the literature. I have put the value that is suspected to be in error in italics. In these cases, the values that are probably erroneous involved the researchers using gravimetric methods alone to determine the concentrations of their MPPIX stock solutions. Other work using independent analysis of the metal concentrations gave different values; however, it is still unclear how to account for the difference between the CoPPIX values reported in [86] and [87], which differ by about 30% – far more than we would expect for even a crude method.

Table A.1: Extinction Coefficients of various Metalloprotoporphyrin IXs

Metalloporphyrin	Solvent	Wavelength (nm)	ε (mM^{-1}cm^{-1})	Ref.
Fe(III)PPIX	0.1 M NaOH (aq.)	385	58.44	[88]
	—	384	*51*	[89]
	100% DMSO	405	183	[68]
	—	406	*170*	[90]
	Glacial acetic acid	383	85	[89]
Fe(II)PPIX	25% Pyridine, 0.1 M NaOH	557, 540	34.53, 9.62	[77, 79]
	—	557	*32.0*	[76]
	0.1 M NaOH	390	39.6	[89]
Co(III)PPIX	33% Pyridine, 0.01 M NaOH,	569, 536, 425	12.0, 11.5, 132	[87]
	and 0.03 M $K_2Fe(CN)_6$	569, 535, 424	*17, 16, 180*	[86]
	0.1 M Tris-acetate pH 7.9	416	93 ± 10	[54]
Ni(II)PPIX	100% DMSO	400, 522, 560	$\varepsilon_\alpha/\varepsilon_\beta = 2.65$	[91]
	0.1 M Tris-acetate pH 7.9	385	52 ± 5	[54]
Zn(II)PPIX	NaOH (aq.)	412, 544, 582	87.4, 6.17, 6.01	[92]
	Pyridine	425, 551, 588	126, 12.3, 10.0	
	ethyl acetate	415	291	[93]
MnPPIX	0.1 M NaOH (aq.)	462	25	[94]
Sn(III)PPIX	0.5% Pyridine, dilute NH_4OH	406	164	[90]
Ga(III)PPIX	DMSO	413	249	
Cu(II)PPIX	0.1 M Tris-acetate pH 7.9	388	64 ± 5	[54]

Table A.2: Extinction coefficients and peak positions of DGCR8 homologues found to associate with heme at the δ, Soret, and α/β peaks. Coefficients were determined using the pyridine hemochromagen assay; see **Apendix A.2** for details. The α/β peak is a broad peak centered around 560 nm; see Figure 1.5. The *Soret* and δ peaks make up the 'split-Soret' peaks characteristic of DGCR8. Data will be published as part of Senturia et al [85].

	ε_{δ}, λ (nm)	ε_{Soret}, λ (nm)	$\varepsilon_{\alpha/\beta}$, 556 nm
Human NC1	62 ± 2 , 366	74 ± 2 , 450	14.1 ± 0.4
Human HBD	60 ± 1 , 366	74 ± 1 , 450	14.2 ± 0.3
Frog HBD	62 ± 3 , 366	72 ± 3 , 451	14.0 ± 0.6
Bat star HBD	59 ± 3 , 367	70 ± 4 , 450	12.9 ± 0.6

Table A.3: Other Extinction Coefficients of Biochemical Interest

Chromaphore	solvent	wavelength (nm)	ε (mM^{-1}cm^{-1})	Ref.
AMP	(aq.)	260, 259	15.02, 15.04	[95]
CMP	—	260, 271	7.07, 8.74	—
GMP	—	260, 252	12.08, 14.09	—
TMP	—	260, 267	8.56, 9.49	—
UMP	—	260, 262	9.66, 9.78	—
Cystine	6 M GdHCl (aq.)	280	0.125	[96]
Trp	—	280	5.450	—
Tyr	—	280	1.265	—

A.5 Lab Recipes

SEC buffer pH 8.0 (1 L)

400 mM NaCl (23.4 g)

20 mM Tris pH 8.0 (2.4 g ; add HCl to pH 8.0)

SP buffer A pH 8.0 (1 L)

100 mM NaCl (5.8 g)

20 mM Tris pH 8.0 (2.4 g ; add HCl to pH 8.0)

SP buffer B pH 8.0 (1 L)

2 M NaCl (116.9 g)

20 mM Tris pH 8.0 (2.4 g ; add HCl to pH 8.0)

TB medium (1 L)

12 g Tryptone

24 g Yeast Extract

4 mL glycerol

dilute to 900 mL; autoclave

Add 100 mL 10X TB Phosphate

10X TB Phosphate, pH 7.2 (1 L)

23.1 g KH_2PO_4

125.4 g K_2HPO_4 Filter or Autoclave.

T7 Transcription Buffer (10 mL)

4 mL 1 M Tris pH 8.0

2.5 mL 1 M $MgCl_2$

0.4 mL 1 M DTT

0.54 mL 371.8 mM Spermidine

2.56 mL H_2O

Sinapinic Acid solution for MALDI

70% (v/v) Acetonitrile

0.1% TFA

Saturate with sinnapinic acid (solid)

1X SDS Running Buffer (5 L)

30 g Tris Base

142.6 g Glycine

5 g SDS

dilute to 5 L with water.

10X TBE (1 L)

108 g Tris base

55 g Boric Acid

40 mL 0.5 M EDTA, pH 8.0

Should be ≈ pH 8.3

Coomasie G-250 stain

30% (v/v) Ethanol

10% (v/v) glacial acetic acid

0.2 % (w/v) Coomasie G250

2X SDS loading dye

2 mL 0.5 M Tris pH 6.8

2 mL 16% SDS

4 mL Glycerol

1 mL 0.2% (w/v) Bromophenol Blue

Dilute to 10 mL with water.

1X PBS (1 L)

8 g NaCl

0.2 g KCl

1.44 g Na_2HPO_4

0.24 g KH_2PO_4

Prescission Protease Buffer

50 mM Tris pH 7.0

150 mM NaCl

1 mM EDTA

1 mM DTT

10X TH(E) buffer (1 L)

340 mM Tris (41.2 g)

660 mM HEPES (157.3 g)

1 mM EDTA (optional)

SEC pH 8.0 plus 8 M Urea (1 L)

480.5 g urea

23.4 g NaCl

2.4 g Tris base; adjust pH to 8.0 with HCl

REFERENCES

[1] BARTEL, D. P. (2009) *Cell*, *136*, pp. 215–233.

[2] GRIFFITHS-JONES, S. (2006) *Methods Mol Biol*, *342*, pp. 129–138.

[3] BORCHERT, G. M., LANIER, W. AND DAVIDSON, B. L. (2006) *Nat Struct Mol Biol*, *13*, pp. 1097–1101.

[4] LEE, Y., KIM, M., HAN, J., YEOM, K.-H. *et al.* (2004) *EMBO J*, *23*, pp. 4051–4060.

[5] CAI, X., HAGEDORN, C. H. AND CULLEN, B. R. (2004) *RNA*, *10*, pp. 1957–1966.

[6] KOZLOWSKI, P., STAREGA-ROSLAN, J., LEGACZ, M., MAGNUS, M. AND KRZYZOSIAK, W. J. (2008) In YING, S.-Y. (ed.), *Current Perspectives in microRNAs (miRNA)*, pp. 1–16, Springer Netherlands.

[7] CHAULK, S. G., THEDE, G. L., KENT, O. A., XU, Z. *et al.* (2011) *RNA Biol*, *8*, pp. 1105–1114.

[8] FALLER, M. AND GUO, F. (2008) *Biochim Biophys Acta*, *1779*, pp. 663–667.

[9] CENIK, E. S. AND ZAMORE, P. D. (2011) *Curr Biol*, *21*, pp. R446–R449.

[10] FORTIN, K. R., NICHOLSON, R. H. AND NICHOLSON, A. W. (2002) *BMC Genomics*, *3*, p. 26.

[11] WU, H., XU, H., MIRAGLIA, L. J. AND CROOKE, S. T. (2000) *J Biol Chem*, *275*, pp. 36957–36965.

[12] LEE, Y., AHN, C., HAN, J., CHOI, H. *et al.* (2003) *Nature*, *425*, pp. 415–419.

[13] DENLI, A. M., TOPS, B. B. J., PLASTERK, R. H. A., KETTING, R. F. AND HANNON, G. J. (2004) *Nature*, *432*, pp. 231–235.

[14] LANDTHALER, M., YALCIN, A. AND TUSCHL, T. (2004) *Curr Biol*, *14*, pp. 2162–2167.

[15] HAN, J., LEE, Y., YEOM, K.-H., KIM, Y.-K., JIN, H. AND KIM, V. N. (2004) *Genes Dev*, *18*, pp. 3016–3027.

[16] GREGORY, R. I., YAN, K.-P., AMUTHAN, G., CHENDRIMADA, T. *et al.* (2004) *Nature*, *432*, pp. 235–240.

[17] SHIOHAMA, A., SASAKI, T., NODA, S., MINOSHIMA, S. AND SHIMIZU, N. (2003) *Biochem Biophys Res Commun*, *304*, pp. 184–190.

[18] STARK, K. L., XU, B., BAGCHI, A., LAI, W.-S. *et al.* (2008) *Nat Genet*, *40*, pp. 751–760.

[19] SCHOFIELD, C. M., HSU, R., BARKER, A. J., GERTZ, C. C., BLELLOCH, R. AND ULLIAN, E. M. (2011) *Neural Dev*, *6*, p. 11.

[20] FÉNELON, K., MUKAI, J., XU, B., HSU, P.-K. *et al.* (2011) *Proc Natl Acad Sci U S A*, *108*, pp. 4447–4452.

[21] SHIOHAMA, A., SASAKI, T., NODA, S., MINOSHIMA, S. AND SHIMIZU, N. (2007) *Exp Cell Res*, *313*, pp. 4196–4207.

[22] ZENG, Y., YI, R. AND CULLEN, B. R. (2005) *EMBO J*, *24*, pp. 138–148.

[23] ZENG, Y. AND CULLEN, B. R. (2005) *J Biol Chem*, *280*, pp. 27595–27603.

[24] HAN, J., LEE, Y., YEOM, K.-H., NAM, J.-W. *et al.* (2006) *Cell*, *125*, pp. 887–901.

[25] FALLER, M., TOSO, D., MATSUNAGA, M., ATANASOV, I. *et al.* (2010) *RNA*, *16*, pp. 1570–1583.

[26] SOHN, S. Y., BAE, W. J., KIM, J. J., YEOM, K. H., KIM, V. N. AND CHO, Y. (2007) *Nat. Struct. Mol. Biol.*, *14*, pp. 847–53.

[27] SENTURIA, R., FALLER, M., YIN, S., LOO, J. A. *et al.* (2010) *Protein Sci*, *19*, pp. 1354–1365.

[28] FALLER, M., MATSUNAGA, M., YIN, S., LOO, J. A. AND GUO, F. (2007) *Nat. Struct. Mol. Biol.*, *14*, pp. 23–29.

[29] TEICH, M. AND NEEDHAM, D. M. (1992) *A Documentary History of Biochemistry, 1770-1940*. Fairleigh Dickinson University Press, Rutherford, NJ, USA.

[30] BERTINI, I., GRAY, H. B. AND VALENTINE, J. (2007) *Biological Inorganic Chemistry: Structure and Reactivity*. University Science Books, Sausalito, CA, USA.

[31] STOKES, G. G. (1863) *Proceedings of the Royal Society of London*, *13*, pp. 355–364.

[32] VIERORDT, K. (1873) *Die Anwendung des Spectralapparates zur Photometric der Absorptionsspectren und zur quantitativen chemischen Analyse.* Laupp, Tübingen.

[33] SEVERINGHAUS, J. W. AND ASTRUP, P. B. (1986) *J Clin Monit*, *2*, pp. 270–288.

[34] DRABKIN, D. L. (1927) *Journal of Biological Chemistry*, *75*, pp. 443–479.

[35] FLINK, E. B. AND WATSON, C. J. (1942) *Journal of Biological Chemistry*, *146*, pp. 171–178.

[36] KEILIN, D. (1925) *Proceedings of the Royal Society of London. Series B.*, *98*, pp. 312–339.

[37] STERN, K. G. (1935) *Nature*, *136*, p. 335.

[38] STERN, K. G. (1936) *J Biol Chem*, *114*, pp. 473–494.

[39] PEISACH, J., BLUMBERG, W. E. AND ADLER, A. (1973) *Ann N Y Acad Sci*, *206*, pp. 310–327.

[40] MORE, C., BELLE, V., ASSO, M., FOURNEL, A. *et al.* (1999) *Biospectroscopy*, *5*, pp. S3–18.

[41] HIRSCH, R. E. (2002) In LAKOWICZ, J. R., GEDDES, C. D. AND LAKOWICZ, J. R. (eds.), *Topics in Fluorescence Spectroscopy*, volume 6, pp. 221–255, Springer US.

[42] VICKERY, L. E. (1978) *Methods Enzymol*, *54*, pp. 284–302.

[43] GRAY, H. B. AND WINKLER, J. R. (1996) *Annu Rev Biochem*, *65*, pp. 537–61.

[44] GRAY, H. B. AND WINKLER, J. R. (2010) *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, *1797*, pp. 1563 – 1572.

[45] SZUNDI, I., CAPPUCCIO, J. AND EINARSDÓTTIR, O. (2004) *Biochemistry*, *43*, pp. 15746–15758.

[46] YEOM, K.-H., LEE, Y., HAN, J., SUH, M. R. AND KIM, V. N. (2006) *Nucleic Acids Res*, *34*, pp. 4622–4629.

[47] BARR, I., SMITH, A. T., SENTURIA, R., CHEN, Y. *et al.* (2011) *J Biol Chem*, *286*, pp. 16716–16725.

[48] BARR, I., SMITH, A. T., CHEN, Y., SENTURIA, R., BURSTYN, J. N. AND GUO, F. (2012) *Proc Natl Acad Sci U S A*, *109*, pp. 1919–1924.

[49] GONG, M., CHEN, Y., SENTURIA, R., ULGHERAIT, M., FALLER, M. AND GUO, F. (2012) *Protein Science*, *21*, pp. 797–808.

[50] FISCHER, H. AND ZEILE, K. (1929) *Annalen der Chemie*, *468*, pp. 98–116.

[51] FALK, J. (1975) *Porphyrins and Metalloporphyrins*. Elsevier Science Ltd, New York.

[52] KADISH, K., SMITH, K. M. AND GUILARD, R. (eds.) (1999) *The Porphyrin Handbook*, volume 1–10. Academic Press, New York, USA.

[53] SMITH, A. T., MAJTAN, T., FREEMAN, K. M., SU, Y., KRAUS, J. P. AND BURSTYN, J. N. (2011) *Inorg Chem*, *50*, pp. 4417–4427.

[54] OZOLS, J. AND STRITTMATTER, P. (1964) *J Biol Chem*, *239*, pp. 1018–1023.

[55] HORI, H., TSUBAKI, M., YU, N. T. AND YONETANI, T. (1991) *Biochim Biophys Acta*, *1077*, pp. 392–399.

[56] KIKUCHI, G., YOSHIDA, T. AND NOGUCHI, M. (2005) *Biochem Biophys Res Commun*, *338*, pp. 558–567.

[57] SHAN, Y., LAMBRECHT, R. W., DONOHUE, S. E. AND BONKOVSKY, H. L. (2006) *FASEB J*, *20*, pp. 2651–2653.

[58] VALLIER, H. A., RODGERS, P. A. AND STEVENSON, D. K. (1993) *Life Sci*, *52*, pp. PL79–PL84.

[59] TANFORD, C. (1962) *Journal of the American Chemical Society*, *84*, pp. 4240–4247.

[60] TANFORD, C. (1964) *Journal of the American Chemical Society*, *86*, pp. 2050–2059.

[61] PACE, C. N. (1986) *Methods Enzymol*, *131*, pp. 266–280.

[62] HARGROVE, M. S., BARRICK, D. AND OLSON, J. S. (1996) *Biochemistry*, *35*, pp. 11293–11299.

[63] CÁRDENAS, M. L. AND CORNISH-BOWDEN, A. (1993) *Biochem J*, *292*, pp. 37–40.

[64] FROST, A. A. AND PEARSON, R. G. (1961) *Kinetics and Mechanism: a Study of Homogenous Chemical Reactions.* Wiley, New York, 2nd edition.

[65] CORNISH-BOWDEN, A. (2012) *Fundamentals of Enzyme Kinetics*. Portland Press Limited, Weinheim, Germany, 4th edition.

[66] MAEHLY, A. C. AND ÅKESON, A. (1958) *Acta Chem Scand*, *12*, pp. 1259–1273.

[67] YANAGI, Y., SEKUZU, I., ORII, Y. AND OKUNUKI, K. (1972) *J Biochem*, *71*, pp. 47–56.

[68] COLLIER, G. S., PRATT, J. M., DE WET, C. R. AND TSHABALALA, C. F. (1979) *Biochem J*, *179*, pp. 281–9.

[69] KUŽELOVÁ, K., MRHALOVÁ, M. AND HRKAL, Z. (1997) *Biochimica et Biophysica Acta (BBA) - General Subjects*, *1336*, pp. 497 – 501.

[70] HILL, A. V. (1910) *Journal of Physiology*, *40 (Suppl.)*, pp. iv–vii.

[71] WEISS, J. N. (1997) *FASEB J*, *11*, pp. 835–41.

[72] CORNISH-BOWDEN, A. AND KOSHLAND, JR, D. E. (1975) *J Mol Biol*, *95*, pp. 201–212.

[73] ANSON, M. L. AND MIRSKY, A. E. (1928) *J Gen Physiol*, *12*, pp. 273–288.

[74] EDSALL, J. T. (1972) *Journal of the History of Biology*, *5*, pp. 205–257.

[75] HILL, R. (1929) *Proceedings of the Royal Society of London. Series B, Containing Papers of a Biological Character*, *105*, pp. 112–130.

[76] DE DUVE, C. (1948) *Acta Chem Scand*, *2*, pp. 264–289.

[77] PAUL, K. G., THEORELL, H. AND ÅKESON, A. (1953) *Acta Chem Scand*, *7*, pp. 1284–1287.

[78] SINCLAIR, P. R., GORMAN, N. AND JACOBS, J. M. (2001) *Curr Protoc Toxicol*, *Chapter 8.3*.

[79] BERRY, E. A. AND TRUMPOWER, B. L. (1987) *Anal Biochem*, *161*, pp. 1–15.

[80] JCGM (2008) *Evaluation of Measurement Data — Guide to the Expression of Uncertainty in Measurement*, volume 100. Joint Committee for Guides in Metrology, Geneva, 1st edition.

[81] MANDEL, J. (1964) *The Statistical Analysis of Experimental Data*. Dover, New York, 1st edition.

[82] TAYLOR, J. R. (1997) *An Introduction to Error Analysis: the Study of Uncertainties in Physical Measurements*. University Science Books, Sausalito, Calif., 2nd edition.

[83] GLENN, A. L. (1960) *J Pharm Pharmacol*, *12*, pp. 595–608.

[84] EDWARDS, C. H. (1994) *Advanced Calculus of Several Variables*. Dover Publications, New York.

[85] SENTURIA, R., LAGANOWSKY, A., BARR, I., SCHEIDEMANTLE, B. D. AND GUO, F. (2012) *PLoS ONE*, *in press*.

[86] YONETANI, T., YAMAMOTO, H. AND WOODROW, 3RD, G. V. (1974) *J Biol Chem*, *249*, pp. 682–90.

[87] WATKINS, S., BARON, J. AND TEPHLY, T. R. (1980) *Biochem Pharmacol*, *29*, pp. 2319–23.

[88] DAWSON, R. M. C. (1986) *Data for Biochemical Research*. Clarendon Press, Oxford, 3rd edition.

[89] KEILIN, J. (1955) *Biochem J*, *59*, pp. 571–579.

[90] RISH, K. R., SWARTZLANDER, R., SADIKOT, T. N., BERRIDGE, M. V. AND SMITH, A. (2007) *Biochim Biophys Acta*, *1767*, pp. 1107–17.

[91] FINDSEN, E. W., SHELNUTT, J. A. AND ONDRIAS, M. R. (1988) *J Phys Chem*, *92*, pp. 307–314.

[92] LEONARD, J. J., YONETANI, T. AND CALLIS, J. B. (1974) *Biochemistry*, *13*, pp. 1460–4.

[93] PRETLOW, T. P. AND SHERMAN, F. (1967) *Biochimica et Biophysica Acta (BBA) - General Subjects*, *148*, pp. 629 – 644.

[94] YONETANI, T. AND ASAKURA, T. (1968) *J Biol Chem*, *243*, pp. 3996–8.

[95] CAVALUZZI, M. J. AND BORER, P. N. (2004) *Nucleic Acids Res*, *32*, p. e13.

[96] PACE, C. N., VAJDOS, F., FEE, L., GRIMSLEY, G. AND GRAY, T. (1995) *Protein Sci*, *4*, pp. 2411–23.