Reading biological processes from nucleotide sequences

Anand Murugan

A Dissertation

Presented to the Faculty

of Princeton University

in Candidacy for the Degree

of Doctor of Philosophy

Recommended for Acceptance

by the Department of

Physics

Adviser: Curtis G. Callan, Jr.

September 2012

Abstract

Cellular processes have traditionally been investigated by techniques of imaging and biochemical analysis of the molecules involved. The recent rapid progress in our ability to manipulate and read nucleic acid sequences gives us direct access to the genetic information that directs and constrains biological processes. While sequence data is being used widely to investigate genotype-phenotype relationships and population structure, here we use sequencing to understand biophysical mechanisms. We present work on two different systems. First in chapter 2, we characterize the stochastic genetic editing mechanism that produces diverse T-cell receptors in the human immune system. We do this by inferring statistical distributions of the underlying biochemical events that generate T-cell receptor coding sequences from the statistics of the observed sequences. This inferred model quantitatively describes the potential repertoire of T-cell receptors that can be produced by an individual, providing insight into its potential diversity and the probability of generation of any specific T-cell receptor.

Then in chapter 3, we present work on understanding the functioning of regulatory DNA sequences in both prokaryotes and eukaryotes. Here we use experiments that measure the transcriptional activity of large libraries of mutagenized promoters and enhancers and infer models of the sequence-function relationship from this data. For the bacterial promoter, we infer a physically motivated 'thermodynamic' model of the interaction of DNA-binding proteins and RNA polymerase determining the transcription rate of the downstream gene. For the eukaryotic enhancers, we infer heuristic models of the sequence-function relationship and use these models to find synthetic enhancer sequences that optimize inducibility of expression. Both projects demonstrate the utility of sequence information in conjunction with sophisticated statistical inference techniques for dissecting underlying biophysical mechanisms.

Acknowledgments

I want to start by thanking my adviser Curt Callan for five years of unconditional support in my research activities. I greatly appreciate the freedom and independence I was given while he guided me toward the completion of interesting projects. While my productivity waxed and waned over the years, I credit him for letting me find my footing naturally, in a new field.

None of the work presented in this thesis would exist if not for the great set of colleagues and collaborators that I have been lucky to work with. Justin Kinney was instrumental in the genesis and execution of my first project in biophysics and I thank him for giving me the opportunity. I would also like to thank Aleksandra Walczak, Thierry Mora and Tarjei Mikkelsen for the fruitful collaborations we have had. I learned a lot from all of them.

I'm grateful for the general biophysics community that surrounded me at the Lewis-Sigler Institute. In particular, I thank Prof. Bialek for providing advice and support whenever needed and for broadening my understanding of what biophysics might be. The students and postdocs I saw everyday over the last few years were also crucial in my education. I thank Yi Deng, XinXin Du, Tiberiu Tesileanu, Dima Krotov, Julien Dubuis, Audrey Sederberg, Gordon Berman, Thibaud Taillefumier, Miriam Osterfield and everyone else who often answered my technical questions, listened to me vent and generally provided great social support and conversation at tea time.

While I have met a large number of exceptional people in my time at Princeton, I am lucky to have befriended a subset that are also nice. I thank Arijeet, Tibi, XinXin, Richard, Pablo, Alvaro, Lucas, Alex, John and Guillaume for being great friends; Hans and Miroslav for being such interesting characters; Rodolfo for putting up with my 'ideas', prejudices and monologues; and Ruggero, Greg, Ida and Krystal for being awesome housemates. I will miss all of them.

One of the hardest things about these five years has been being 3000 miles from friends that I value greatly. I thank Alysia and Stephen for being such awesome friends and Mario for being Mario. While I didn't see enough of you, remembering that you guys exist was always comforting.

I have to acknowledge Ravishankar, my high school physics teacher and friend, who showed me the joy of thinking for myself. It's hard to exaggerate how much our discussions about physics, inquiry and understanding meant to me.

Finally, and most importantly, I thank my family for making any of this possible. Arvind and Ingrid, thanks for all your help (and the dinners). Amma and Appa, thank you for your love and support. I don't know how you figured it out but you are brilliant parents.

Contents

Chapter 1

Introduction

Nucleic acids are the most fundamental of biomolecules. They contain the instructions for assembly of all the molecular components necessary for the functions of life (including their own replication); themselves serve crucial roles in the assembly process; and additionally contain instructions for and play active roles in the highly complex regulation of the production of various biological molecules.

The ongoing revolution in our ability to read nucleic acids [1] has allowed the investigation of many biological questions that were previously inaccessible. In particular, great progress has been made in cataloguing genes and non-coding functional elements within and across species; revealing evolutionary history and population structure of species; and in associating genetic variation with medically relevant phenotypes [2].

Sequence data can also provide insight into fundamental processes of molecular biology. Examples include the discovery of RNA interference [3], transcriptional regulation by distant DNA sequences [4, 5], recombination hotspots [6], and numerous insights into biochemical pathways [7]. However, sequence data also provides the opportunity for quantitative characterization of biological systems.

In this thesis, we present work on two different biological processes where we use high-throughput sequence data to quantitatively understand the molecular processes

that underly the design of functional DNA sequences. First, in chapter 2, we focus on the remarkable process of VDJ recombination that stochastically edits germline DNA in precursor T-cells to produce diverse T-cell receptors in the human immune system. We infer statistical distributions of the underlying biochemical events that generate T-cell receptor coding sequences from the statistics of the observed sequences.

These distributions characterize the potential repertoire of T-cell receptors that can be produced by an individual. Since the observed repertoire of T-cell receptors is a product of selective forces from exposure to antigen as well as the molecular constraints of VDJ recombination, our model of the latter serves as a baseline for analysis of the effect of selection on the repertoire. Additionally, we are also able to calculate the generation probability of any specific T-cell receptor, including as yet unobserved ones.

Then, in chapter 3, we present work on modeling the function of regulatory DNA sequences in both prokaryotes and eukaryotes. We use measurements of the transcriptional activity of large libraries of mutagenized promoters and enhancers to infer models of the sequence-function relationship. For the well-studied *lac* promoter, we infer the parameters of a physically motivated 'thermodynamic' model of the transcription rate of the regulated gene. This model accounts for the interaction of DNA-binding proteins and RNA polymerase with each other and with the sequence. Eukaryotic enhancers are more complex in mechanism, with larger number of proteins that interact at overlapping binding sites. For these, we infer heuristic models of the sequence-function relationship and use these models to synthesize enhancer sequences that optimize the function of the enhancer in desired ways.

Chapter 2

Generation of T-cell receptor diversity

2.1 Introduction

This chapter is an expanded and edited version of a manuscript submitted for publication. The work described was done in collaboration with Aleksandra Walczak, Thierry Mora and Curtis Callan. The data analyzed was contributed by Harlan Robins.

All organisms are threatened by pathogens and hence have defense mechanisms to protect against them. The fundamental problem to be solved by these mechanisms is to distinguish between self and non-self biological material and destroy the latter. There are two classes of such mechanisms: those that act generally against all pathogens (innate) and those that are specifically adapted to an invading pathogen (adaptive).

The innate immune system is the more ancient of the two and such defenses are found in all species. Bacteria use restriction enzymes to cleave foreign DNA; plants and animals all produce antimicrobial peptides and use Toll-like receptors to induce

their production upon (non-specific) recognition of foreign pathogens [8]; colonies of unicellular eukaryotes show phagocytosis where some individual cells sacrifice themselves by ingesting bacteria [9]; and all vertebrates and many invertebrates have 'complement systems' that make phagocytosis more efficient [10].

An adaptive response to pathogens requires a mechanism for variation in the design of receptors that recognize pathogens. Adaptive immune systems were until recently believed to exist only among jawed vertebrates, but have recently been discovered in jawless vertebrates as well [11] – likely a case of convergent evolution since the implementations are very different in the two. Thus all vertebrates have specialized cells carrying antigen receptors that undergo somatic diversification.

The generation of diverse antigen receptors requires stochastic editing of genetic information that codes for these receptors. This is implemented as a remarkable process called V(D)J recombination [8, 12], where specific (V,D and J) genes are chosen randomly from the germline DNA and joined together to produce a new surface receptor protein each time a new immune system cell is generated. In the beta chain of human T-cell receptors (the focus of this work) the germline has 48 different V-genes, 2 D-genes and 13 J-genes (each having a few alleles), so that the number of possible VDJ combinations is only a few thousand, far too small to account for the full diversity of the immune system. The bulk of this diversity comes instead from a process in which, during separate DJ and VD joining events, a random number of bases are deleted from the ends of the two genes being joined and a variable length of random sequence is inserted between them. The end product is the so-called CDR3 region of the receptor gene: a short, highly variable region that plays an essential role in determining the antigen specificity of the cell.

In this chapter, we focus on characterizing the statistics of V(D)J recombination in humans using sequencing of entire TCR repertoires. The next sections provide a basic overview of the human adaptive immune system and the biochemistry of

receptor generation. The subsequent sections describe the sequence data we use to study this system and the analysis of this data.

2.2 Adaptive immune system in humans

2.2.1 Lymphocytes

The adaptive immune system in humans consists of T-cells and B-cells, both derived from the same pluripotent hematopoietic stem cells. While T-cells are formed from the stem cells in the thymus, B-cell differentiation occurs in the bone marrow. Both types of lymphocytes carry receptors on their surface that recognize foreign antigen. B-cell receptors (BCRs) are membrane bound forms of antibodies which are secreted by the cells upon activation. BCRs recognize antigen found in the exterior of cells. T-cell receptors (TCRs), on the other hand, are designed to recognize peptide fragments from foreign pathogens within cells, displayed on the cell surface by the host cells. The peptides are presented by a membrane glycoprotein complex called the major histocompatibility complex (MHC).

T-cells and MHCs each come in two varieties. Cytotoxic ($CD8^+$) T-cells recognize peptides bound to MHC class I molecules which display peptides from proteins in the cytosol, thus potentially signalling viral infections. Helper ($CD4^+$) T-cells recognize peptides bound to MHC class II molecules which are found only on antigen presenting cells (APCs) – macrophages, dendritic cells and B-cells – that have internalized bacteria and display peptides from proteins in their vesicles. While cytotoxic T-cells kill the infected cells upon recognition, helper T-cells activate the APCs causing them to proliferate. The CD8 and CD4 co-receptors aid each TCR in the process of antigen recognition. The TCRs in cytotoxic and helper T-cells are generated identically. Upon recognition of an MHC bound peptide, T-cells themselves are activated becoming a 'memory' T-cell (as opposed to their initial 'naive' state) and also proliferate.

2.2.2 T-cell receptor architecture and formation

Each T-cell has about 30,000 identical receptors on its surface [8], coded for in the genome of that cell by a DNA sequence that has been stochastically edited during differentiation from the stem cell. The structure of an $\alpha:\beta$ T-cell receptor and the genes that are combined to produce it are shown schematically in Fig. 2.1. A minority $(1-10\%)$ of peripheral T-cells carry a different kind of TCR, called the $\gamma:\delta$ receptor, whose function is not yet clear. We will focus purely on $\alpha:\beta$ T-cells.

Two polypeptide chains, α and β, each containing a constant domain (lower; blue) and a variable domain (upper; red, green and yellow) make up the receptor. The genes that code for all of these domains are spread on chromosome 7 over 1000 kilobases and 620 kilobases for the α and β chains respectively [13]. The antigen binding site on the TCR is composed of six hypervariable loops (see crystal structure in Fig. 2.1), three each from the α and β chains, that are brought together spatially in the TCR, to form three distinct complementarity determining regions (CDR1, CDR2 and CDR3). The corresponding CDR nucleotide sequences in the α and β chains are highly variable.

Of these, CDR3 shows the highest variability on both chains because it is the junctional site of the remarkable process of V(D)J recombination [14] which occurs in precursor T-cells in the thymus. In this process genes that are far apart on the genome are cut, brought together and joined, accompanied by random deletion and insertion of nucleotides. The CDR3 of the α chain is formed by the joining of one V_α gene (of 70 choices) and one J_α gene (of 61 choices). The variable domain of the β chain is formed by the joining of one V_β gene (of 48 choices), one D_β gene (of 2 choices) and one J_β gene (of 13 choices). The CDR1 and CDR2 loops are completely contained within the V_α and V_β genes and their diversity comes from the different choices of these genes. The focus of the analysis in this chapter is the statistics of V(D)J recombination events in the β chain.

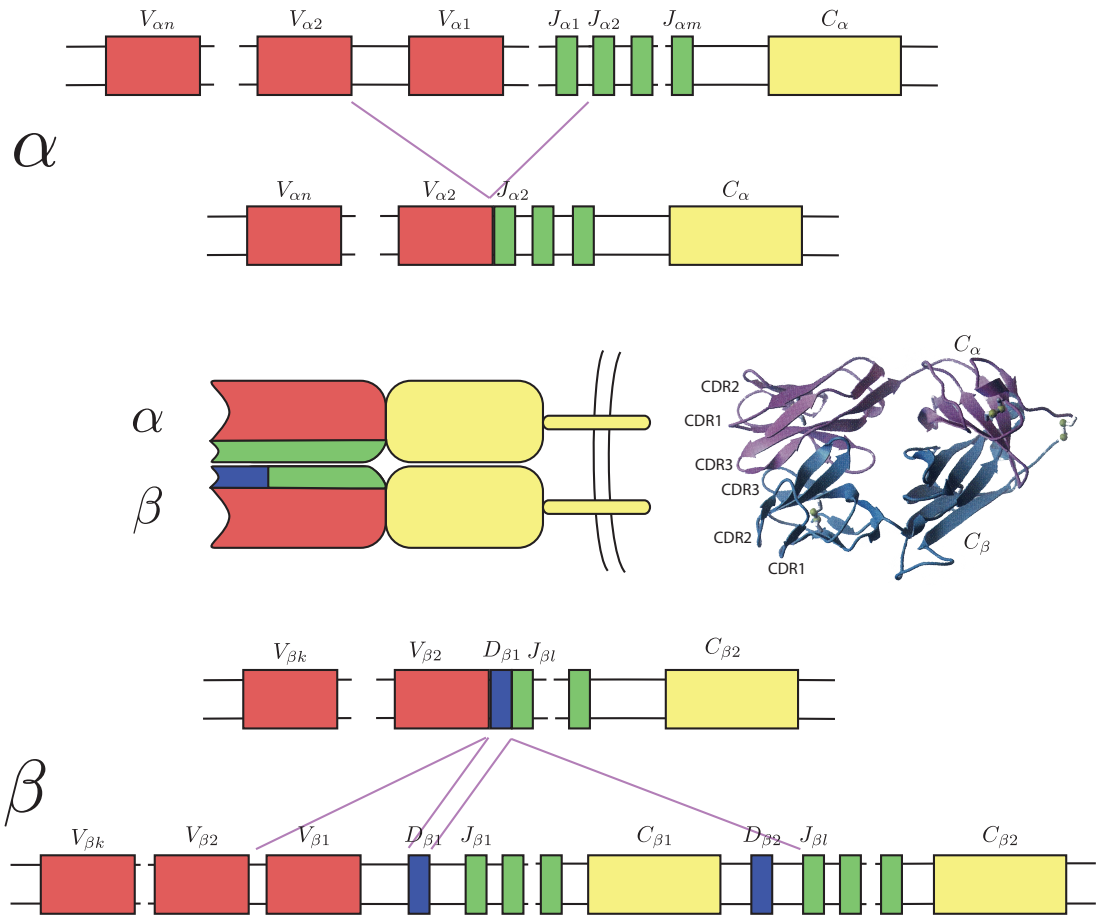

Figure 2.1: T-cell receptor architecture. The α chain is composed by the joining of one V_α gene and one J_α gene while the β chain is composed by two joining events, one D_β gene with one J_β gene and one V_β gene with the $D_\beta - J_\beta$ combination. The 'constant' domains C_α and C_β form the base of the TCR at the cell surface. The antigen binding site has the hypervariable CDR3 region which is the junctional site of these recombination events. The crystal structure of the TCR is shown (figure adapted from [8]).

2.2.3 Clonal and thymic selection

The clonal expansion of lymphocytes when activated by high-affinity binding of antigen serves to preferentially increase the number of lymphocytes that are of utility in the immune response to the identified threat. Crucially, this selective mechanism can only work if each lymphocyte has a single type of receptor on its surface, thus allowing the clonal selection to act on specific receptor shapes [15]. The history of exposure to various antigen thus exerts a strong influence on the repertoire of TCRs found in an individual.

There are also strong selective forces that act on newly differentiated T-cells. Since TCRs that bind self protein peptides are highly undesirable, new T-cells are subject to a battery of tests in the thymus. T-cells with receptors showing high affinity to self peptides are eliminated by apoptosis. Additionally, TCRs that show very weak affinity to the tested peptides are also unlikely to be of use to the immune system, and hence these are eliminated as well. Thus there is both positive and negative selection for the antigen recognition potential of new TCRs. In fact, $\sim 98\%$ of thymocytes die in the thymus [8].

2.2.4 Biochemistry of V(D)J recombination

VDJ recombination is implemented by a set of DNA processing enzymes that act in a complex series of steps and proceeds by first recombining D with J, then V with DJ. In the first step, the recombination activating gene (RAG) protein complex, directed by recognition signal sequences (RSS) flanking the genes, brings two randomly chosen D- and J-genes together, cuts out the intervening chromosomal DNA, and forms a hairpin loop at the end of each gene [16, 17]. In further steps [18, 19] the hairpin loops are opened, creating overhangs at the end of both genes that may eventually survive as P-nucleotides (short inverted repeats of gene terminal sequence) [20]. This is followed by nucleotide deletions and insertions that can be template-dependent or

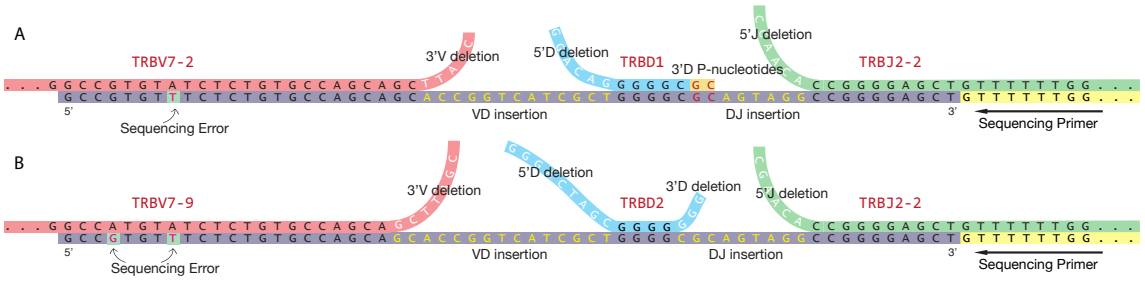

Figure 2.2: A 60bp CDR3 read (grey box) can be aligned to different genes (nomenclature follows IMGT conventions [22]) with different deletions (white), insertions (yellow), and P-nucleotides (red). (A) Alignment to specific V-, D-, and J-genes with insVD=13, insDJ=6, delV=5, delJ=6, del5'D=6, del3'D=−2 (in other words, pal3'D=2). (B) Alignment of the same read to different V- and D-genes, and with insVD=15, insDJ=9,delV=7, del5'D=9, del3'D=3 (no P-nucleotides). Note that the alignment to the V-gene is not maximal in this case. A few heavily penalized mismatches are allowed (in the V-gene in this example) in order to accommodate a small sequencing error rate. The location of the sequencing primer is indicated: it is chosen to uniquely identify the start of the CDR3 read within each J-gene.

simply random. After ligation, the end result is a variable coding joint between the chosen D- and J-genes [21], and the whole process is repeated to make another coding joint between a randomly chosen V-gene and the outcome of the D-J joining process.

2.3 Analysis strategy and sequence data

2.3.1 Sequence data from immune cell receptor repertoires

We work with sequence data on CD4+ T-cell beta chain CDR3 regions obtained from nine human subjects as described in [23]. In these experiments, T-cells are collected from a blood sample, and sorted into 'naïve' (CD45RO-; cells that have not been activated by recognition of antigen) and 'memory' (CD45RO+; cells that have been activated by recognition of antigen) compartments, DNA is extracted, and sequence reads long enough to capture a $5'$ piece of the J gene, a $3'$ piece of the V gene and the variable sequence lying in between, are obtained. Each sequence is read

multiple times, and a clustering algorithm is used to correct for sequencing error [23] (see section 2.6.5 for more information on errors). This process produces a data set consisting of an average of 232,000 (140,000) unique CDR3 sequences from the naïve (memory) compartments for each individual subject[1]. Each unique sequence comes with a multiplicity (ranging over three orders of magnitude) reflecting the prevalence of that particular cell type in the blood sample.

2.3.2 Isolating molecular constraints from selection

As described, the repertoire of TCRs in the blood is shaped by the molecular constraints from V(D)J recombination as well as clonal and thymic selection. Since we are here interested in characterizing the outcomes of the recombination process pre-selection, we focus on a special subset of rearranged CDR3 nucleotide sequences.

A majority of recombination events do not result in a productive TCR. The random insertion and deletion of nucleotides can shift the genes out of the correct reading frame, or result in a premature stop codon. When such a non-productive rearrangement occurs, a second rearrangement attempt can be made on the other chromosome in the cell. If this attempt is successful and the cell survives thymic selection, we now have a cell with two rearranged CDR3s, one of which is non-productive. This subset of non-productive CDR3 sequences has not been subject to selection, either thymic or clonal. Therefore their statistics purely represents the molecular constraints of V(D)J recombination.

Roughly 14% of the unique CDR3 sequences are non-productive. We focus our analysis on these non-productive CDR3 sequences, of which there are an average of 35,000 (22,000) in the naïve (memory) compartments for each individual subject. We analyze the naïve and memory data sets separately.

[1]We are grateful to H. Robins and collaborators for making the data sets on which this work is based available to us.

2.3.3 Recombination events from nucleotide sequences

Each recombined sequence can be thought of as the outcome of a generative event described by several random variables (Fig. 2.2): V-, D-, and J-gene choices, deletions of variable numbers of nucleotides from the selected genes, insertions of random nucleotides between them, and the possible creation of P-nucleotides (short palindromic nucleotides at the end of the gene segments as in Fig. 2.2A at the $3'$ end of the D-gene). The statistical distribution of these event variables in a population of newly-created receptors is an important quantity: it contains information about the *in vivo* functioning of the biochemical editing mechanism and provides the baseline for a quantitative assessment of the downstream workings of selection in the adaptive immune system.

We wish to infer this distribution from the large T-cell sequence repertoires that are becoming available via high-throughput sequencing technology [24–26]. To date, this inference has been done via a deterministic alignment procedure which assigns a unique event to each sequence [24, 25]. However, since individual CDR3 sequences can arise in multiple ways (see Fig. 2.2), this assignment really should be done on a probabilistic basis; this is particularly true since, as we will show, deterministic alignment introduces spurious correlations in the statistics of generative events (Fig. 2). In short, a formal statistical inference procedure is needed to accurately infer the underlying event probability distribution from the data. We present such a method, based on likelihood maximization via an iterative expectation-maximization algorithm [27], and apply it to recent data on human T-cell receptor sequences.

2.3.4 Structure of recombination event distributions

Each CDR3 generating recombination event can be fully characterized by a set E of discrete variables:

$$E_{CDR3}: \begin{cases} \text{V, D, J} \\ \text{del}V\text{, del}J\text{, del}D5\text{, del}D3 \\ \text{pal}V\text{, pal}J\text{, pal}D5\text{, pal}D3 \\ \text{ins}VD\text{, ins}DJ \\ (x_1, \ldots, x_{insVD}),\ (y_1, \ldots, y_{insDJ}) \end{cases} \tag{2.1}$$

These variables comprise the identities of the V-, D- and J-genes selected for recombination[2] (V,D,J); the numbers of bases deleted from the 3' end of the V-gene (del$V$), the 5' end of the J-gene (del$J$), and both ends of the D-gene (del$5'D$ and del$3'D$ for the 5' and 3' ends, respectively); the number of palindromic nucleotides at each of the gene ends ($\text{pal}V, \text{pal}J, \text{pal}5'D, \text{pal}3'D$); the specific sequence $(x_1, \ldots, x_{\text{ins}VD})$ of length insVD inserted at the VD junction, and the specific sequence, $(y_1, \ldots, y_{\text{ins}DJ})$ of length insDJ inserted at the DJ junction (see Fig. 2.2). We choose a convention in which both sequences are read in the 5' to 3' direction, but the VD (DJ) inserted sequence is read from the sense (antisense) strand.

We seek a joint distribution over all of these variables containing the minimal set of dependences between the variables that is required to self-consistently capture the observed correlations in the data. We find that the following factorized form for the probability of a recombination event E (defined by specific values for all the event variables) successfully captures all the significant correlations between sequence features that are present in the data (see Fig. 2.5):

[2]Here we distinguish only the genes, not their various alleles. The gene list includes germline pseudo-genes: they cannot produce functioning receptor proteins but, since we work with non-coding VDJ rearrangements, pseudogene sequences can appear in the data.

$$
\begin{aligned}
P_{\text{recomb}}(E) &= P(V)\,P(D,J)\times \\
&\quad P(\text{del}V|V)\,P(\text{del}J|J)\,P(\text{del}5'D,\text{del}3'D|D)\times \\
&P(\text{ins}VD)\prod_{i=1}^{\text{ins}VD} p_{VD}^{(2)}(x_i|x_{i-1})\,P(\text{ins}DJ)\prod_{i=1}^{\text{ins}DJ} p_{DJ}^{(2)}(y_i|y_{i-1}).
\end{aligned}
\tag{2.2}
$$

The various factors are normalized joint or conditional distributions on their respective arguments. $P(V)$ and $P(D,J)$ account for the fact that the various genes have different usage probabilities (and that D- and J-gene usage is correlated). The factors $P(\text{del}V|V)$, etc., are distributions on the number of nucleotide deletions, conditioned on the gene being deleted (deletion profiles turn out to be very gene-dependent). $P(\text{ins}VD)$ and $P(\text{ins}DJ)$ give the probabilities of different numbers of nucleotide insertions at each junction. The parameters $p_{VD}^{(2)}$ and $p_{DJ}^{(2)}$ account for possible nucleotide bias in the insertions: they give the conditional probabilities of inserting a specific nucleotide given the identity of the immediately 5' nucleotide, with x_0 referring to the last nucleotide at the 3' end of the truncated V-gene on the sense strand for a VD insertion, or at the end of the truncated J-gene on the antisense strand for a DJ insertion.

P-nucleotides do not appear explicitly in Eqn. 2.2: we treat them as 'negative' deletions (*i.e.* a palindrome of half-length 2, as in Fig. 2.2A, is counted as a deletion of value -2). This is possible because we find that when the number of nucleotide deletions is greater than zero, occurrences of palindromic nucleotides at the end of the gene segment are completely explained by chance insertions of the corresponding nucleotides (see section 2.4.4). Thus, true P-nucleotides, not attributable to chance insertions, only occur in association with zero nucleotide deletions and it is consistent to label them as 'negative' deletions.

2.3.5 Generation probability and likelihood of observed sequences

The probability $P_{\rm gen}(\sigma)$ of generating a specific CDR3 sequence σ is the sum of the probabilities of all recombination events E_σ that produce σ:

$$P_{\rm gen}(\sigma) = \sum_{E \in E_\sigma} P_{\rm recomb}(E). \tag{2.3}$$

The likelihood $L(\sigma)$ of observing a specific CDR3 sequence read σ, however, must take into account residual sequencing error as well as allelic variation, and is given by a sum over a larger set of recombination events $\widetilde{E}_\sigma$ that generate sequences close to σ:

$$L(\sigma) = \sum_{E \in \widetilde{E}_\sigma} P(E, \sigma) \qquad \text{where} \tag{2.4}$$

$$P(E,\sigma) = P_{\rm recomb}(E) \times \frac{1}{(1+R)^L} \times \sum_{\text{alleles } a} P(V_{\rm a}|V_E) P(J_{\rm a}|J_E) P(D_{\rm a}|D_E) \left(\frac{R}{3}\right)^{n_{\rm err}(\sigma_E^a, \sigma)}. \tag{2.5}$$

In the latter equation, $n_{\rm err}$ is the number of mismatches between the observed read σ and the CDR3 sequence σ_E^a that would be produced by the recombination event E with allele choices a. L is the length of the sequence read. The mismatch rate R is determined in the inference with the rest of the distribution parameters and reflects both sequencing error as well as unknown allelic variation. In practice, we only consider recombination events $\widetilde{E}_\sigma$ that lead to CDR3 sequences with at most a few mismatches from σ. The sum over alleles[3] arises because we do not know *a priori* which alleles are present and reads may not go deep enough into the gene sequence to clearly distinguish alleles from each other [28]. The probabilities of the different

[3]We use the known alleles for each gene listed in the IMGT data base [22] augmented by a few additional variants observed in the data (see Appendix section 2.6.1 for details).

alleles, given a gene, are also inferred and are expected to differ from individual to individual.

The likelihood of the whole data set $\mathcal{D}$ is then the product over the individual sequence likelihoods: $\mathcal{L}(\mathcal{D}) = \prod_{\sigma \in \mathcal{D}} L(\sigma)$. This expression depends implicitly on the parameters defining the generative probability distribution (along with the allele distributions and the sequencing error parameter), and we infer their correct values by maximizing $\mathcal{L}(\mathcal{D})$. In order to identify universal features of the diversity generation machinery, we perform this inference separately for each individual subject.

2.3.6 The expectation maximization algorithm

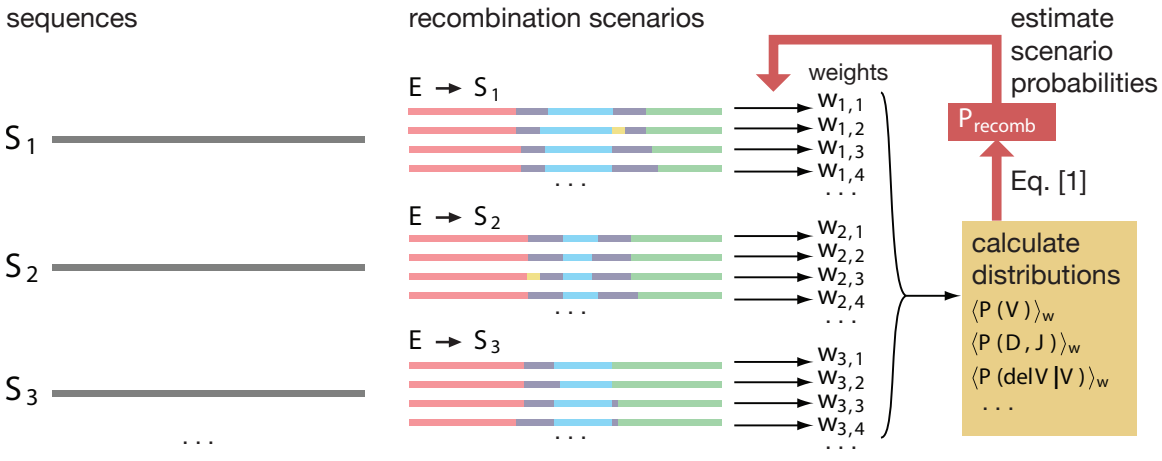

Figure 2.3: Flow chart of the analysis pipeline.

There are two major steps in the analysis pipeline that leads from a list of CDR3 sequences to a final estimate of the probability distribution $P_{\text{recomb}}(E)$ of generative recombination events. The first is an 'alignment' step in which, for each read σ, we create a comprehensive list of recombination 'scenarios' $\{E_\sigma\}$ that could plausibly have produced that read. A 'scenario' is a particular set of values for the event variables (gene identities, VD insertions, etc.) that generates a recombined sequence nearly identical to the read in question (with possibly a small number of mismatches).

This step is described in greater detail in section 2.6.3. The second major step is an expectation maximization algorithm (summarized in the flow chart of Fig. 2.3) for finding the generative distribution that maximizes the likelihood of the observed data given the functional form of the generative distribution (as expressed in Eqn. 2.2).

We wish to find model parameters that maximize the likelihood of the data. We use an iterative Expectation-Maximization algorithm [27,29] to do this. Given a current guess for the model parameters that describe $P_{\rm recomb}(E)$, we update it by calculating the probability-weighted counts of events over the data set and then using those counts to re-estimate the marginal distributions ($P(V)$, $P(D,J)$, $P(\mathrm{ins}VD)$, and so on) that appear as factors in the general functional form of $P_{\rm recomb}(E)$ (Eqn. 2.2).

As indicated in Eqns. 2.3-2.5, the joint likelihood of a recombination event E and sequence σ is the product of two factors: the probability of the generative event (given by $P_{\rm recomb}(E)$), and the sum over allele choices a of the probability of those allele choices multiplied by the probability of the number of mismatches between σ and the sequence σ_E^a implied by E and a. In other words, in addition to the recombination event probability $P_{\rm recomb}(E)$, likelihood involves the sequencing error rate R and the allele probabilities $P(V_a|V)$, etc. We emphasize that we carry out this exercise independently for the data sets derived from different individuals. While we expect (and find) that $P_{\rm recomb}(E)$ is consistent between individuals, we of course expect different individuals to have different allele probabilities.

In the expectation maximization procedure, we start from a prior in which each factor in Eqn. 2.2 for $P_{\rm recomb}(E)$ is uniform in its variables, the sequencing error rate R is set to a small value (typically 10^{-4}), and the allele probabilities are uniform over all the alleles of each gene. Using Eqn. 2.5, for each CDR3 sequence read σ, we exhaustively compute the likelihoods of all recombination events E given σ, starting from maximal alignments for each sequence identified in the initial parsing of the read (previous section), and looping over the other scenarios, involving extra deletions

compensated by chance re-insertions of identical nucleotides, that could also 'explain' the read. We also loop over the number of true P-nucleotides in the cases where they are present.

Normalizing these likelihoods yields the relative weights that observing the sequence σ assigns to different recombination events E, given the current model parameters. Summing these weighted occurrences over all the sequences in the data set gives a new, data-conditioned, estimate of the various factors that enter into the assumed general form of $P_{\text{recomb}}(E)$ (as well as a new estimate of the sequencing error probability and allele occurrence frequencies). The formal statement of the update rule is as follows; for each parameter in the model that describes the probability of a specific recombination event feature X (say a particular V-gene choice) we update it to the probability weighted counts over the whole data set of that event. In other words, the $(k+1)$-th iteration of the model parameters are given by

$$\begin{aligned} P^{(k+1)}(X) &= \sum_{\sigma\in\mathcal{D}} \sum_{E} \delta_{X_E,X}\, P^{(k)}(E|\sigma) \\ &= \sum_{\sigma\in\mathcal{D}} \sum_{E} \delta_{X_E,X}\, \frac{P^{(k)}(E,\sigma)}{L^{(k)}(\sigma)} \end{aligned} \tag{2.6}$$

where $\delta_{X_E,X}$ is one if X is true in the recombination event E and zero otherwise. This procedure is used to update all the factors entering into the likelihood calculation and the process is repeated until convergence to a stable end point is achieved. Since all sequences in the data set are looped over in the calculation, we can record 'on the fly' the likelihood $L(\sigma)$, the generation probability $P_{\text{gen}}(\sigma)$ of that sequence (a conceptually different quantity), as well as the conditional entropy of events $S(E|\sigma)$ for each sequence quantifying the multiplicity of recombination events that could have produced the given CDR3 sequence). The product of $L(\sigma)$ over all sequences is the current overall likelihood of the data set, a measure of convergence of the procedure. The generation probabilities $P_{\text{gen}}(\sigma)$ have a direct physical significance, reflecting the

probability of generation of the sequence by the molecular machinery.

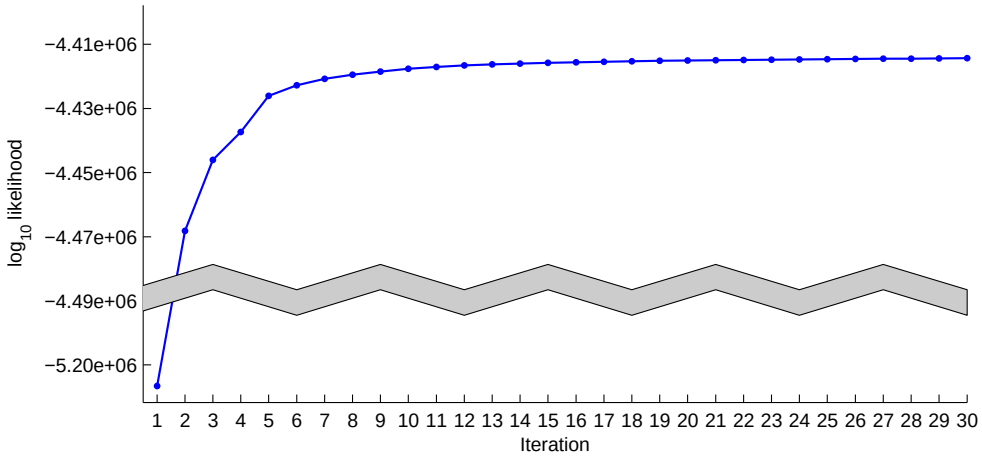

Figure 2.4: Convergence of the total likelihood of all data sets with iterations of the EM algorithm.

Iterating this process is guaranteed, by general expectation maximization arguments, to maximize the overall likelihood of the data set locally. We have found that rapid and direct convergence to a likelihood maximum is the norm for the data sets we work with (see Fig. 2.4). The models for the probability distribution of generative events inferred in this way from the different data sets are available online[4].

2.4 Results

The factors in our equation for $P_{\text{recomb}}(E)$ (Eqn. 2.2) are probability distributions on event variables that take on a finite number of values. The joint distribution has a total of 2865 parameters (more than 90% of which are needed to specify the deletion profiles of the individual V-, D- and J-genes). Despite the large number of parameters, we are able to determine them accurately and without overfitting. We emphasize that our goal is to obtain an accurate description of recombination event statistics, and not

[4] `www.princeton.edu/~ccallan/TCRPaper/Models`

(yet) to explain those statistics mechanistically. In what follows, we present results of our analysis of naïve, non-productive, CDR3 sequence repertoires of nine individuals (see section 2.4.9 for the results of analysis of memory sequence repertoires).

2.4.1 Correlations between event variables

It is important to verify that correlations not present in the assumed structure of the probability distribution (Eqn. 2.2) are in fact not present in the data. To perform this self-consistency check, we use the inferred generative distribution to compute the probability-weighted counts distribution of recombination event variables in the data, and then use this distribution to calculate the mutual information of all pairs of event variables. The matrix of mutual information values is shown in the upper-triangular part of Fig. 2.5A, where the entries outlined in red are dependences accounted for by individual factors in our assumed form of $P_{\text{recomb}}(E)$ (Eqn. 2.2), entries outlined in green are indirect dependences that can be induced by these factors, and the rest would vanish if the data were perfectly described by the assumed structure of $P_{\text{recomb}}(E)$. There are a few detectable correlations that are not consistent with the assumed structure: $(\text{ins}VD, \text{del}V), (\text{ins}DJ, \text{del}J)$ and (V, D). They are, however, all so weak (mutual information < 0.02 bits) that we do not model them explicitly (indeed, they might arise from subtle biases in our inference procedure).

For comparison, in the lower-triangular part of Fig. 2.5A we show the mutual information values of all pairs of variables, but now calculated from a deterministic assignment of events to sequences based on maximal alignments. The resulting distributions exhibit spurious correlations that are absent from the corrected, maximum likelihood estimate (MLE) of the distributions. For instance, the number of insertions at the two junctions are found to be independent in our analysis while the uncorrected estimate shows a dependence (Fig. 2.5B,C).

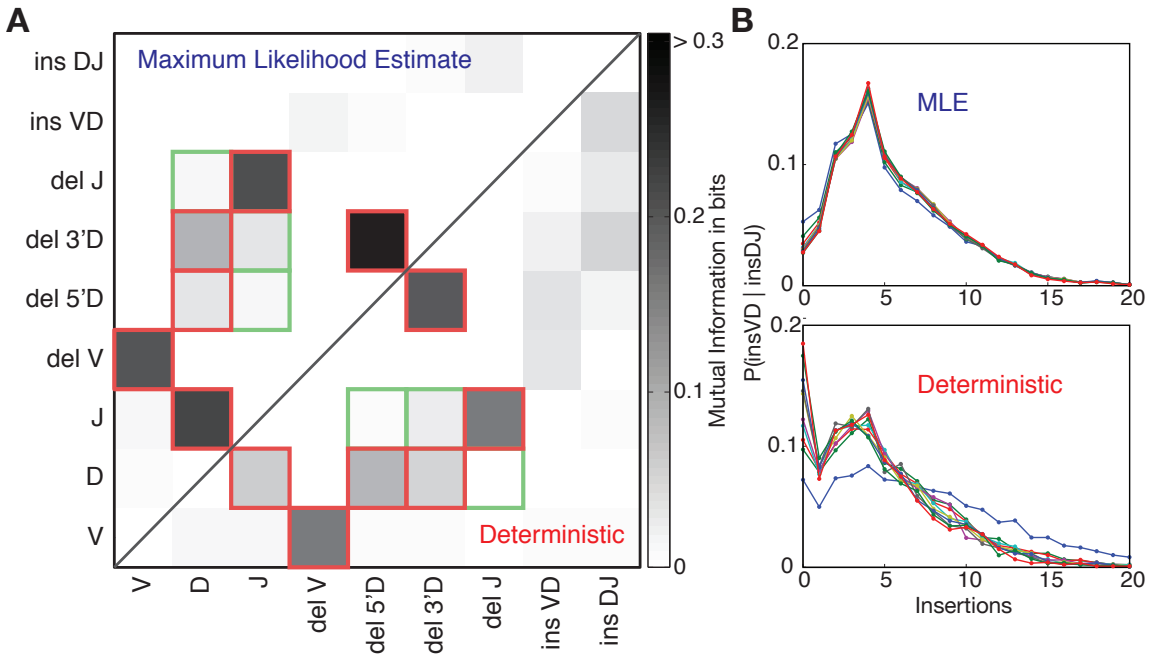

Figure 2.5: (A) Data-derived correlations between sequence features: each entry is the mutual information $I(X, Y)$ of a feature pair over the naïve non-productive repertoire. The outlined elements are correlations expected from the form of $P_{\text{recomb}}(E)$: red identifies a direct effect of a factor in Eqn. 2.2 (e.g. D $\leftrightarrow$ J) and green indirect effects (e.g. D $\leftrightarrow$ J $\leftrightarrow$ delJ). The top-left half of the matrix shows results from the maximum likelihood estimate (MLE), while the bottom-right half corresponds to a deterministic maximum-alignment based identification of recombination events. (B) Probability distribution of the number of VD insertions conditioned on the number of DJ insertions for MLE (top) and deterministic (bottom) analysis. Each curve corresponds to a different value of insDJ, ranging from 0 (blue) to 10. The curves collapse for MLE indicating independence.

2.4.2 Gene usage distributions

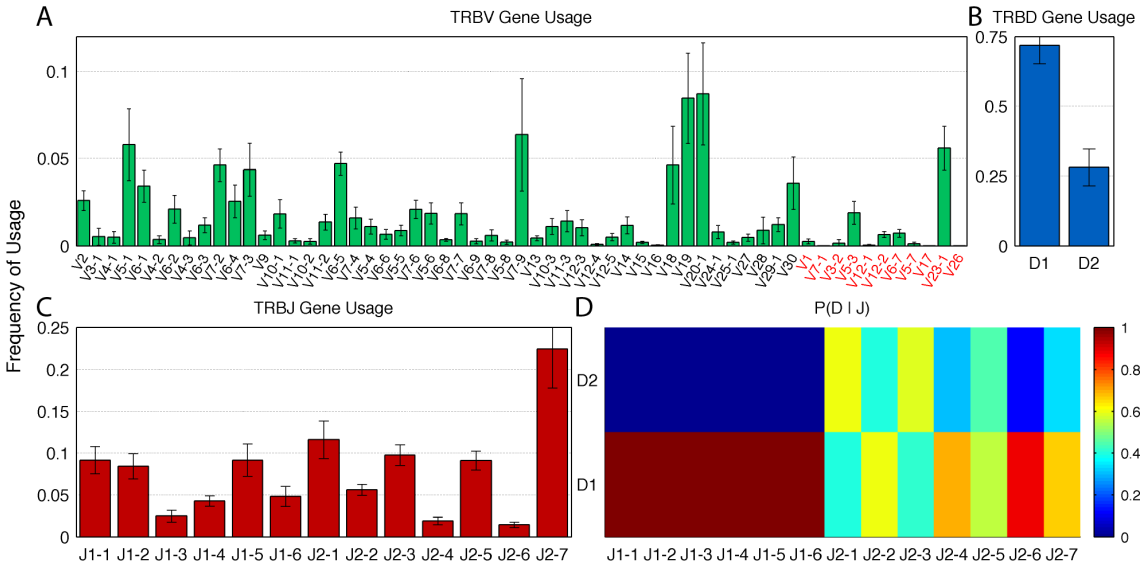

Figure 2.6: Statistical aspects of gene usage. (A) Usage frequencies of V-genes, ordered by position on the chromosome, with the exception of pseudogenes (red legend). (B) Usage frequencies of the two D-genes. (C) Same for the 13 J-genes. (D) D-gene usage frequencies, conditioned on J-gene choice. As expected from the mechanistic constraint, TRBD2 has essentially zero probability ($< 0.1\%$) of recombining with any TRBJ1 gene. Error bars indicate variation across the nine individuals.

In Fig. 2.6, we show the inferred gene usage frequencies. The frequencies of V- and J-genes vary significantly from gene to gene, a phenomenon for which no mechanistic explanation has yet been given. In particular, linear location on the chromosome does not explain the pattern of either V- or J-gene usage (the genes in Fig. 2.6A,C are ordered by position). The usage frequencies are consistent between individuals, though of all the inferred parameters in P_{recomb}, these usage patterns show the most relative variation between individuals.

The pattern of D-gene use conditioned on J-gene choice (Fig. 2.6D) reveals the known mechanistic constraint prohibiting utilization of D-genes that lie $3'$ of the chosen J-gene [8]. The inferred distribution assigns a total probability of less than 0.1% for joining events using TRBD2 and any TRBJ1 gene. We note that such a determina-

tion is impossible without probabilistic analysis due to the uncertainty in identifying genes in specific sequences. The dependence between V gene choice and D or J gene choice is very weak to non-existent (with mutual information less than 0.01 bits). Thus, we believe that previously reported correlations in the use of these genes [30] reflect the effects of selection rather than VDJ recombination.

Finally, we note the presence of pseudo V-genes which occur in almost 10% of the non-productive CDR3s. These pseudogenes cannot produce a functional receptor but they can participate in the recombination process and produce a non-productive rearranged CDR3 sequence which can be transmitted into the naïve or memory compartments just like any other non-productive rearrangement. The set of V gene sequencing primers used by Robins et. al. [26, 31] either exactly or approximately match 11 pseudogenes. Of these, TRBV23-1, TRBV5-3, TRBV12-2 and TRBV6-7 show significant usage.

2.4.3 Nucleotide insertions

In Fig. 2.7 we show the factors related to insertions in the inferred distribution $P_{\text{recomb}}(E)$. The VD and DJ insertions are uncorrelated (Fig. 2.5) and their length distributions are nearly identical, with exponential tails (Fig. 2.7A). The nucleotide frequencies in the inserted segments are not uniform and are well explained by a di-nucleotide Markov model where the probability of inserting A, C, G, or T depends on the immediately 5' nucleotide (see Fig. 2.7B). The VD inserted segment, on the sense strand, and the DJ inserted segment, on the antisense strand, show a preference for Cs. The frequencies of tri-nucleotides are almost perfectly accounted for by the di-nucleotide preferences (Fig. 2.7C), giving evidence that the sequence statistics are fully captured by our inferred dinucleotide statistics. Additionally, the VD insertion di-nucleotide bias, taken on the sense strand in the 5'-3' direction, is virtually identical to the DJ insertion di-nucleotide bias, taken on the antisense strand in the

5'-3' direction. This suggests that the mechanism of junctional nucleotide insertions is strand specific and occurs on opposite strands for the VD and DJ junctions. The molecular mechanistic basis of these features is not evident.

2.4.4 Palindromic nucleotides co-occur only with zero nucleotide deletions

Mechanistically, it is thought that short palindromic nucleotides at the edges of the inserted segments result from the opening of the hairpin ends of the cleaved genes. Thus we expect such true P-nucleotides to co-occur only with zero nucleotide deletions since if nucleotides are actually deleted from the end of the gene, the palindromic nucleotides would also be lost. However, we do expect accidental palindromes from the chance insertion of appropriate nucleotides.

To show that the occurrence of palindromic nucleotides with non-zero nucleotide deletions from the ends of the genes is consistent with chance insertions, we keep track of the (model probability weighted) joint frequencies of lengths of observed palindromes conditioned on the number of deletions and on gene choice. Keeping track of this detail is necessary because of the strong dependence of deletion probabilities on gene choice. After we obtain our converged model, we calculate the frequencies of chance palindromic nucleotides of different lengths co-occurring with non-zero deletions (taking into account all the structure of $P_{\text{recomb}}(E)$, including the nucleotide bias in insertions). The plot in Fig. 2.8 shows that the observed frequencies of palindromic nucleotides co-occurring with non-zero deletions are completely consistent with those expected by chance insertions. Thus, P-nucleotides can be consistently counted as 'negative' deletions as they occur only in association with zero nucleotide deletions.

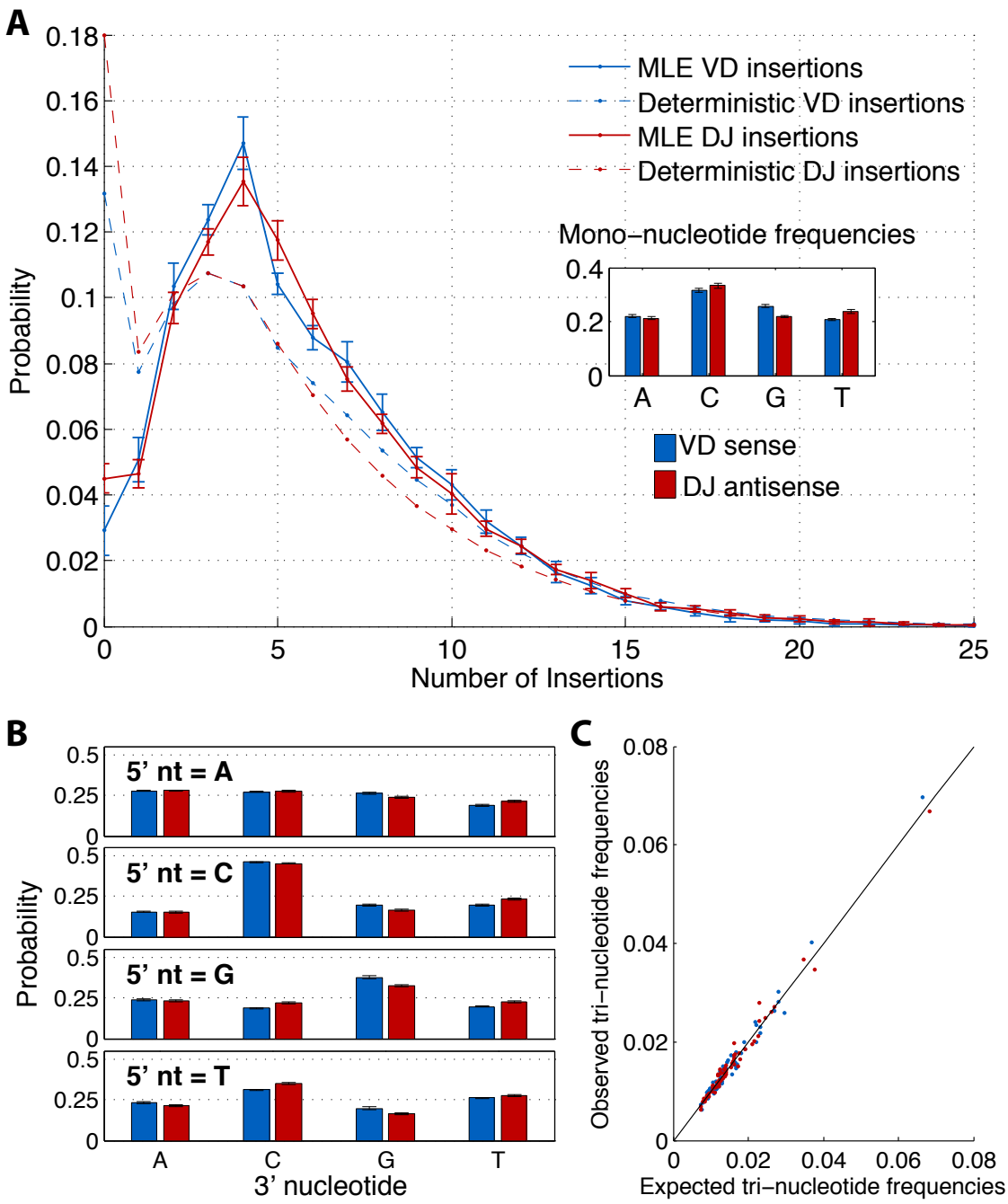

Figure 2.7: Statistics of VD and DJ insertions. (A) Insertion length profiles: maximum likelihood estimate (deterministic estimate) displayed as solid (dashed) lines; error bars show variation across the nine individuals. The distribution tail is accurately exponential. The deterministic estimate greatly overestimates the frequency of zero insertions. Inset: mono-nucleotide utilization bias. (B) Dinucleotide utilization in insertions; the bias in DJ insertions is very accurately the reverse complement of the VD insertion bias.(C) Higher-order nucleotide bias in VD (blue) and DJ (red) insertions is completely accounted for by dinucleotide statistics.

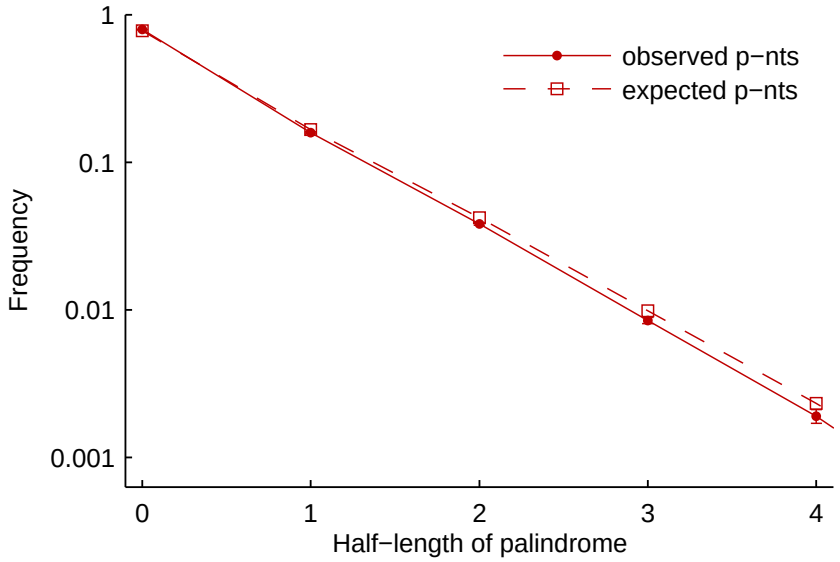

Figure 2.8: Occurrence frequency of P-nucleotides for non-zero deletions.

2.4.5 Nucleotide deletions

Since there is a strong correlation between number of deletions and gene identity (see the entries for $I(\text{del}V, V)$ and $I(\text{del}J, J)$ in Fig. 2.5), we allow for gene-dependent deletion profiles in $P_{\text{recomb}}(E)$ (Eqn. 2.2). The results for a few genes are shown in Fig. 2.9A (see Appendix Figs. 2.16-2.20 in section 2.6.7 for all the profiles). The profiles have substantial variation from gene to gene, suggestive of a nuclease activity that depends on sequence context, but they are highly consistent between individuals. We have modeled this context dependence using a position weight matrix summing independent contributions from the bases in a 6 nucleotide window (four $3'$ and two $5'$) around the cutting point to the log probability of deletion (see Fig. 2.9B and Appendix section 2.6.7 for details). We find that only bases $3'$ of the deletion site have a strong effect on the probability, with T and A nucleotides having the greatest contribution, consistent with previous observations [32]. This simple model, which ignores both the P-nucleotides as well as the effects of distance from the end of the gene, does reasonably well in explaining the variation in deletion probabilities

($r^2 = 0.7$). We believe more sophisticated but parsimonious mechanistic models could reproduce our inferred distributions.

2.4.6 Consistency of distributions across individuals

The distributions of insertions and deletions are highly consistent between individuals (Figs. 2.7,2.9), including the dozens of very different gene-dependent deletion profiles (Fig. 2.9A,C). This suggests that the statistics of the repertoire of unique non-productive, sequences reflect a universal molecular mechanism of rearrangement. In addition, the consistency of the distributions inferred from completely independent data sets is convincing evidence against overfitting. We note, however, that for certain specific inferred probabilities of insertions, deletions and gene choice, the uncertainty from our very large but finite sample size accounts for less than 50% of the observed inter-individual variance (indicated by the error bars in all figures), possibly reflecting biological variation.

2.4.7 Potential diversity of repertoire

Our inferred distribution of recombination events (Eqn. 2.2) implies a probability distribution $P_{\rm gen}(\sigma)$ on the space of all CDR3 sequences (Eqn. 2.3) whose entropy $S_{\rm seq} = -\sum_\sigma P_{\rm gen}(\sigma) \log P_{\rm gen}(\sigma)$ (roughly speaking, the logarithm of the effective number of unique sequences that can be generated) is a measure of the potential sequence diversity of VDJ recombination. Since multiple recombination events can lead to the same sequence, we cannot calculate $S_{\rm seq}$ directly. We do, however, have an explicit description of $P_{\rm recomb}$, the entropy of which we can calculate: $S_{\rm recomb} = 52$ bits; in addition, we can show that sequence entropy and recombination event entropy are related by

$$S_{\rm seq} = S_{\rm recomb} - \langle S(E|\sigma)\rangle_\sigma \simeq 47 \text{ bits}\,, \tag{2.7}$$

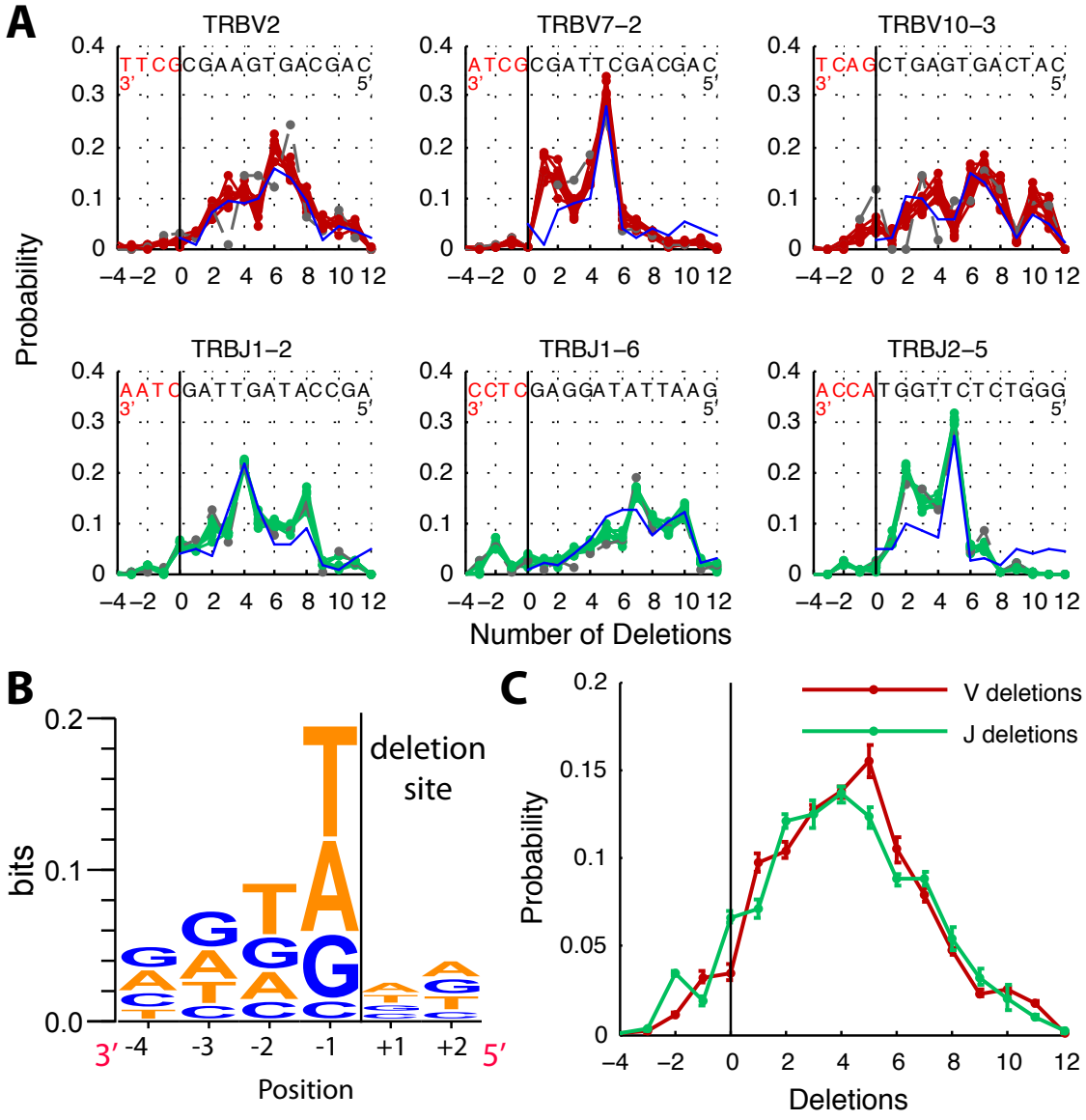

Figure 2.9: (A) Gene-specific deletion profiles for selected V (red) and J (green) genes: the profiles vary widely from gene to gene, but are nearly identical across individuals (all nine are plotted; one in grey from an individual with significantly smaller sample size). The blue curves in all panels show the predictions of a simple model for the sequence context dependence of deletion probabilities using a position weight matrix (PWM), fit to the V deletion profiles (see Appendix section 2.6.7 for details). The model ignores P-nucleotide generation and lacks any effects of distance from the gene end but performs reasonably well ($r^2 = 0.7$). (B) Sequence logo of the context dependence of deletion probability, from the PWM fit to the V deletion profiles. Only positions 3′ of the deletion site have strong effects on the probability. (C) Cumulative deletion profiles for V-genes and J-genes. Error bars indicate variation across individuals.

where the correction term, $\langle S(E|\sigma)\rangle_\sigma \simeq 5$ bits, is the entropy of recombination events that give the same sequence, averaged over sequences. This means that CDR3 sequences can be generated in ~ 32 different ways, on average, by VDJ recombination; this is the fundamental reason why we cannot assign a unique generative event to each individual CDR3 sequence and therefore must resort to probabilistic inference methods. The total sequence diversity of 47 bits corresponds to a potential CDR3 repertoire size of $\sim 10^{14}$ sequences[5]. This is to be compared with the estimated 4×10^6 unique CDR3 sequences in an individual [26] , the $\sim 10^{11}$ T-cells in the blood of an individual [33] and the $\sim 10^{13}$ potential peptide-MHC complexes [34]. While convergent recombination means that the sequence entropy cannot be neatly partitioned into contributions from gene choice, deletions and insertions, the entropy of recombination events S_{recomb} can be so partitioned (Fig. 2.10A). We note that the bulk (60%) of the recombination entropy comes from the nucleotide insertions, and little from gene choice (5 bits from V and 4 bits from D and J) consistent with previous estimates [35]. For comparison, uniform usage of the genes would result in an entropy of 5.9 bits for V and 4.7 bits for D and J gene choices.

2.4.8 Overlap of repertoires between individuals

A striking feature of the data is that there are sequences that appear in the repertoires of more than one individual, and it is interesting to assess whether the shared sequences (both their number and their specific identities) are consistent with chance on the basis of our generative distribution $P_{\text{gen}}(\sigma)$. We see evidence of inter-sample contamination in some of our data leading to a large number of shared sequences between specific individuals. Eliminating such questionable cases (see Appendix section 2.6.6 for details), we are left with 21 sequences that occur in the non-productive repertoires of two individuals and none that occur in more than two.

[5]Recall that this estimate is for the β-chain only. The α-chain will yet add more diversity to this estimate.

The number of shared sequences between the repertoire samples of any pair of individuals with sample sizes N_1 and N_2 is expected to be Poisson distributed with mean $\bar{n} = N_1 N_2 \langle P_{\text{gen}} \rangle_\sigma$ where $\langle P_{\text{gen}} \rangle_\sigma = \sum_\sigma P_{\text{gen}}^2(\sigma) \simeq 3.4 \pm 0.1 \times 10^{-10}$ is the average value of the probability of producing a CDR3 sequence σ, estimated by taking the mean of P_{gen} over the obserbed repertoire. In Fig. 2.10B, we compare the expected number of pairs of individuals with a certain number of shared sequences (calculated as a sum of Poisson distributions over the pairs) to the observed number of such pairs, showing excellent agreement. The specific shared sequences have particularly high generation probabilities according to our distribution, with a median value of $\sim 10^{-8}$ compared to the repertoire median of $\sim 10^{-14}$ (Fig. 2.10C). Since the generative distribution is trained on individual repertoires, and is highly consistent between individuals, its success in accounting for recurring sequences between individuals is a non-trivial test of its validity. We find similar results for the shared sequences among the memory repertoires (see Fig. 2.12).

2.4.9 Memory T-cell non-productive repertoire

Inference of $P_{\text{recomb}}(E)$ from the non-productive memory repertoires of the same nine individuals leads to results identical with those reported above for the naïve non-productive repertoires. The non-productive CDR3 sequences in both of these compartments should not be subject to selection. The consistency of the inferred generative distribution between these repertoires as well as between the nine individuals is strong evidence that the non-productive CDR3 sequence statistics, memory or naïve, reflect only the basic recombination process and not selection.

While the naïve non-productive compartments contain an average of 35,000 unique sequences per individual, the memory non-productive compartments are smaller, containing an average of 22,000 unique sequences per individual. In Fig. 2.11, we compare the naive and memory insertions and deletions distributions. In Fig. 2.12 we show that

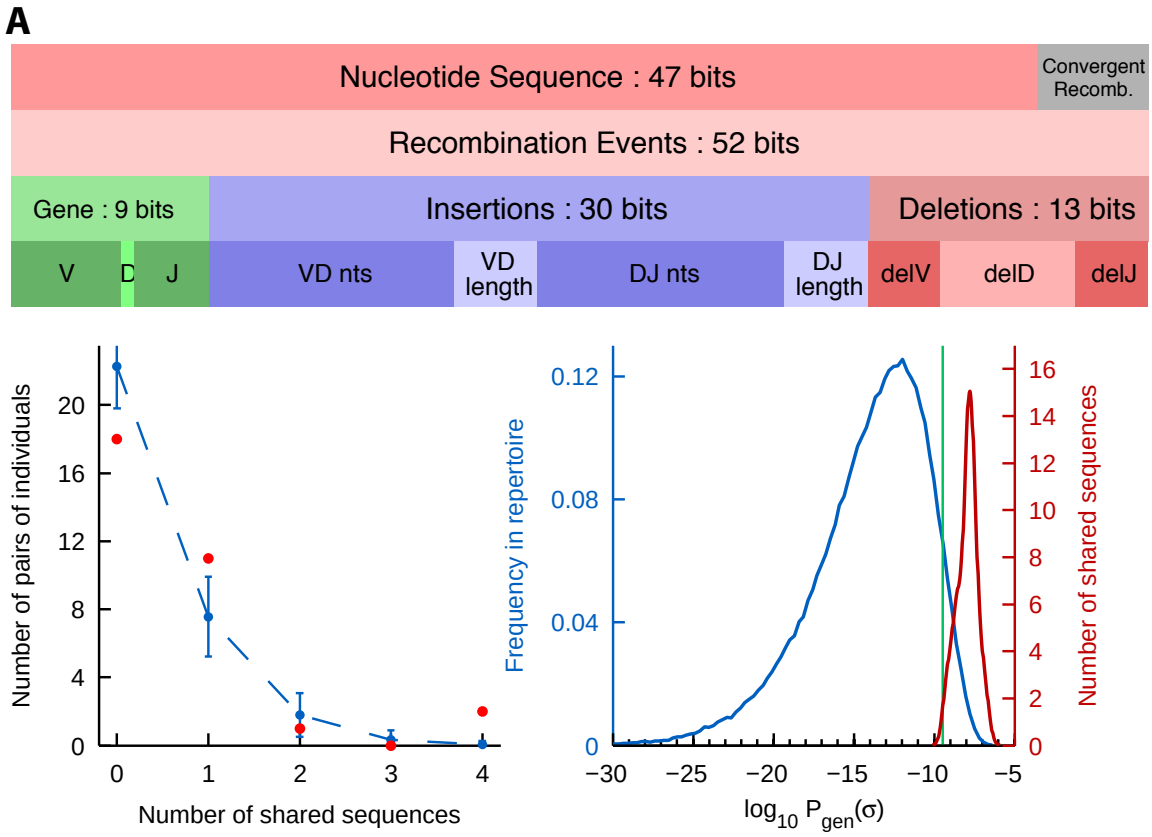

Figure 2.10: (A) Entropy decomposition. Top bars: sequence entropy is smaller than recombination entropy by 5 bits because of convergent recombination; Bottom bars: recombination event entropy decomposed into contributions from gene choice, insertions, and deletions. (B) Statistics of the 21 CDR3 sequences shared between pairs of individuals: actual (red) vs. expected on the basis of the inferred $P_{\rm gen}(\sigma)$ (blue). (C) Histogram of $P_{\rm gen}(\sigma)$ for all sequences (blue) and for the 21 shared sequences (red); $\langle P_{\rm gen}\rangle$ for the full repertoire is indicated by the vertical green line.

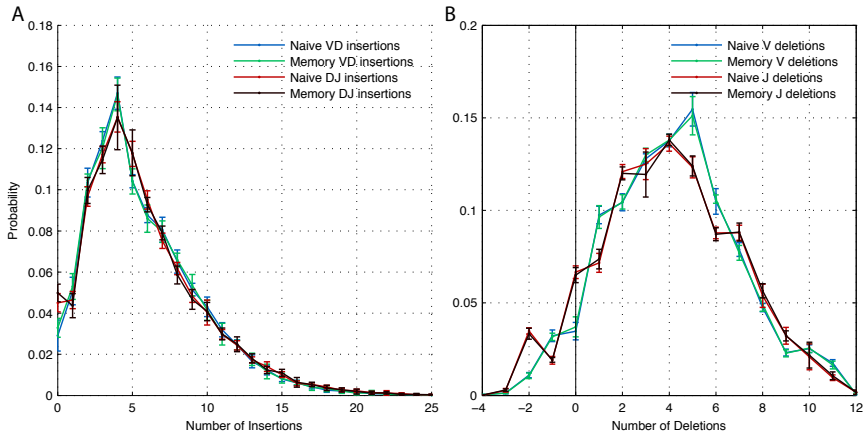

Figure 2.11: Comparison of insertions (A) and deletions (B) distributions for the naive and memory T-cell repertoires. We find that the inferred models from the two compartments are statistically identical in all respects. Error bars indicate variation across the nine individuals.

the occurence of shared sequences between the individual memory non-productive repertoires is consistent with our generative model for the memory compartments as well. The plots show that the models inferred from the naïve and memory T-cells are identical in all respects, in confirmation of the expectation that non-productive sequences are not subject to selection effects.

While it is tempting to apply our approach to the in-frame sequence repertoires, it would be inconsistent to do so: these sequences have passed selection filters, either thymic or adaptive (or both), and we have no analog of Eqn. 2.2 to parametrize the probability that a sequence σ, once produced in a recombination event, will subsequently pass selection filters. This is an important subject for future investigation.

2.4.10 Convergent recombination and generation probability

Convergent recombination, *i.e.* multiple ways of producing the same TCR, has been proposed as an explanation for the occurrence of 'public' TCRs [36–38]. However, the recombination entropy $S(E|\sigma)$ is only weakly correlated with the generation probability $P_{\text{gen}}(\sigma)$ (correlation coefficient 0.13, see Fig. 2.13), and we find that the

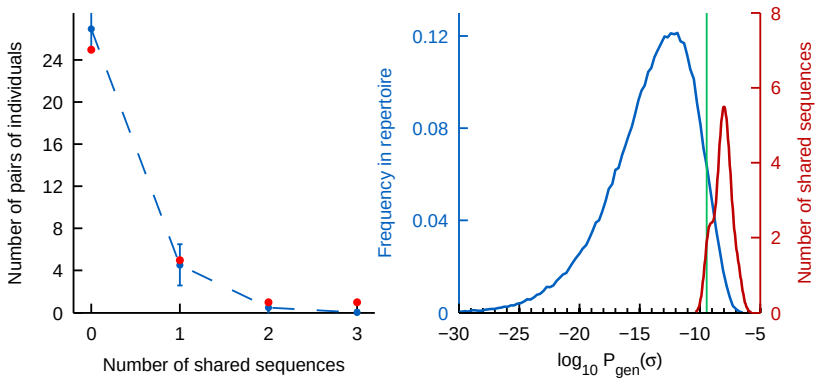

Figure 2.12: Shared sequences in memory T-cell non-productive CDR3 sequence repertoires. A) Distribution of number of shared sequences between the 9 individuals. B) Distribution of $P_{\text{gen}}(\sigma)$ for the entire repertoire (blue) and for the recurring sequences (red). $\langle P_{\text{gen}} \rangle$ is indicated by the green vertical line.

shared non-productive sequences in our data (red dots) do not have notably higher recombination entropies than other sequences.

2.5 Discussion

We have presented a method for inferring the statistics of VDJ recombination events from the large T-cell receptor sequence repertoires that are being made available by high-throughput sequencing. We emphasize the crucial importance of using a probabilistic approach: the typical CDR3 sequence can be produced by about 32 different recombination events, and using a deterministic assignment of events to each sequence results in systematic biases and spurious correlations. Our general approach allows us to cope with not-yet-indexed alleles [28] and, most importantly, with sequencing errors, an essential task given the rapid growth of high-throughput but error-prone sequencing technologies.

Since we focus on non-productive sequences, our results describe the probability distribution over CDR3 sequences produced by the recombination machinery *before any functional selection has occurred*. Its remarkable reproducibility across individ-

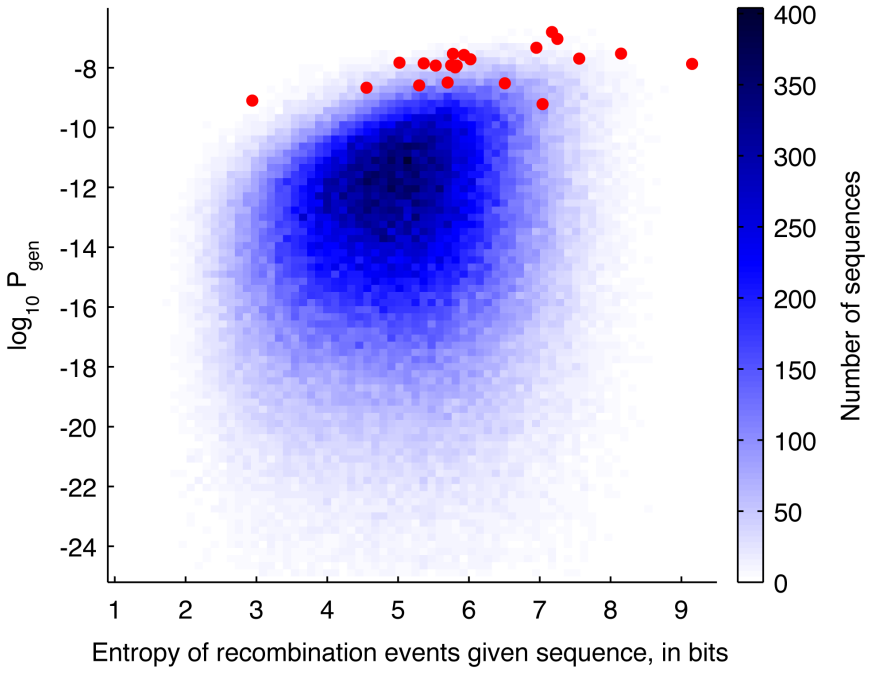

Figure 2.13: A 2D histogram of conditional entropy of recombination events given the sequence and $P_{\text{gen}}(\sigma)$. Convergent recombination (as measureed by the recombination event entropy) is a contributing factor to $P_{\text{gen}}(\sigma)$, with correlation coefficient 0.13. The shared sequences in the naive non-productive repertoires are shown in red.

uals and repertoires (naïve and memory) provides compelling evidence for the consistency and accuracy of our method. The obtained distribution is a central feature of the adaptive immune system and serves as a baseline (or, in evolutionary terms, a neutral model) for analyzing the subsequent processes of the immune system. By calculating the entropy of the generative distribution, we can estimate the potential diversity of the CDR3 sequences ($\sim 10^{14}$ sequences) and the contributions of insertions, deletions and gene choices to this entropy. We find that insertions contribute most (60%) of the diversity.

We are able to evaluate the probability of generating *any* specific CDR3 sequence (including as yet unobserved ones). This probability could be used to estimate the strength of selection on a sequence or group of sequences, or the likelihood that a sequence is shared between individuals or repertoires. Thus, it could help better

characterize the significance of shared or 'public' TCR sequences [38]. We have verified that the sequences that are shared between the non-productive repertoires of different individuals in our data are consistent with the predictions of the inferred probability distribution (Fig. 2.10B,C), a very stringent test of its accuracy.

The recombination event distributions also provide insight into the molecular mechanism of recombination, and should serve as a starting point for detailed mechanistic models of recombination. We find that the recombination processes at the two junctions are essentially independent of each other, and that insertion events are independent of gene choice and deletions. The inferred distribution confirms that a D-gene can only recombine with downstream J-genes. We derive a precise model for the composition of inserted nucleotides, based solely on frequencies of di-nucleotides. We also show that a relatively crude model of sequence-specific nuclease activity can account for the deletion probabilities reasonably well. Our general distribution has a very large number of parameters. Parsimonious, but sufficiently sophisticated, mechanistic models are needed to reproduce the inferred distributions.

We have focused on characterizing the molecular generation of nucleotide sequences that code for T-cell receptors. While the underlying biochemistry conveniently served to parametrize our sequence distributions, finding an analogous functionally relevant parametrization of amino-acid sequences to model the effects of selection is much more challenging [39]. Statistical analysis of the productive receptor repertoires, with our precise characterization of the unselected repertoire in hand, will hopefully aid in this effort.

2.6 Appendix

2.6.1 Sequences of V, D, and J-genes and their alleles

Accurate knowledge of the sequences of germ line V-, D-, and J-genes and their allelic variants is essential to minimize errors and bias in our analysis. There are 2 D-genes, 13 J-genes, and 48 V-genes, not counting alleles. There are in addition 19 'pseudo' V-genes on the same germline chromosome: they participate in the recombination process and, though they cannot lead to a functioning receptor, they can appear in the non-productive sequence data sets, provided that a sequencing primer (or an approximate one) is present, which in our case is true for 11 pseudo V-genes.

We curated a list of known and discovered allelic variants of the V-genes by combining those found in the public IMGT database [22] with variants that we discovered with high confidence during our analysis. Not all the sequence reads listed in IMGT are true variants since many of them are from rearranged DNA with variation at the junctional end. Such 'variants' were removed from our list, unless the variation was deeper in the sequence, far from the edited end. In addition, we have found three instances of allelic variants in our data that are not listed in IMGT. The discovered variants of genes TRBV7-7 and TRBV10-1 can actually be found by BLAST in the NCBI database of human sequences; the variant of gene TRBV7-2 is not found by BLAST and appears to be completely novel. Undiscovered variants have rather small impact on overall recombination event statistics, but they can cause systematic errors in the inference of gene-specific deletion profiles.

Complete lists of the genes and alleles used in our analysis are available online[6]. For completeness, we also list the primers used by Robins et. al. [26, 31] in acquiring the data we analyze.

[6] `www.princeton.edu/~ccallan/TCRPaper/GenData`

2.6.2 CDR3 sequence data files and formats

The CDR3 sequences used in our analysis come from naïve or memory CD4+ T-cells of 9 human individuals, and are further segregated into 'in-frame' and 'non-productive' sequences. The sequences are 60bp in length for 6 of the subjects, and 101bp in length for the remaining three. The reads of different length differ only in how far the sequencing window goes into the V gene: both types are anchored on the same conserved phenylalanine in the J-gene and have the same read depth into the J-gene.

Processed sequence data was made available to us by H. Robins. As described in [26, 31] each sequence is read multiple times and the multiple reads are used to estimate the multiplicity of each specific TCR receptor in its respective compartment. In addition, multiple reads are used to correct for sequencing errors by clustering reads that differ at a small number of positions [31]. In our data files, the effective sequence multiplicity is recorded along with the error-corrected sequence (although we do not use multiplicity in our current analysis). The data files used in our analysis are available online[7]. The file names in the repository clearly indicate the category to which the included data belongs.

2.6.3 Initial parsing of sequence reads by alignment

The first step in our inference procedure is to align each CDR3 read with specific alleles of V, D, and J genes by sequence matching. The goal is to generate a set of plausible recombination events that could produce the read to serve as a starting point for subsequent probabilistic refinement. This preliminary alignment procedure produces, for each read, a finite number of V, D, and J alleles, the maximal length alignments of these alleles to the read, the corresponding minimum nucleotide deletions from the genomic sequences, with possible P-nucleotides identified, and with the unmatched parts of the read identified as VD or DJ insertions. Mismatch information

[7] `www.princeton.edu/~ccallan/TCRPaper/SeqData`

is also stored.

Certain thresholds are imposed on the alignments – gene alignment lengths must be sufficiently long; gene deletions must not be too large; errors are allowed in the alignments (no gaps), but the number of errors must be small. The alignment score (using an appropriate mismatch penalty) is used to rank order alignments, and a threshold on the score relative to the score of the best alignment is also imposed. Specific values for these various parameters are chosen in the light of computational experience to achieve fast and accurate convergence of the overall model-fitting algorithm.

The procedure for finding J matches is simplest. The CDR3 reads all begin at the $3'$ end (sense strand) from a primer in a known position in each J gene. Thus for each candidate J gene, we simply look for exect matches of the end of the sequence read with the portion of the gene just $5'$ of the primer. Proceeding in this way, and imposing the various thresholds mentioned, we find an average of 2-3 J alignments per read.

For the V-gene, the position of alignment to the read is not fixed. So for a given V-gene, we align the $5'$ end of the read to the m-th base from the $3'$ end of the V-gene, and note the best-scoring match at this positioning (this time allowing some mismatches, and penalizing them in the score). We step through the values of m and record the best-scoring match over all positionings. Repeating this process for all the V-genes, and imposing the earlier mentioned thresholds, we are left with a limited set of possible V-gene identifications, together with their specific alignments to the read. Proceeding in this way, we find an average of ~ 15 V alignments per read.

After identifying the plausible alignments to V- and J- genes, we turn to the problem of identifying D-gene matches. This is a more difficult problem because the D-genes are short, and deletions (occurring on both ends) often leave residual sequences which are hard to identify as a D-gene fragment. We therefore put very

loose constraints on the D-gene alignments, relying on the probabilistic refinement to narrow them down. Specifically, we consider the read sequence segment lying between the end of the highest-scoring V-gene and the end of the highest-scoring J-gene, and include 10 nucleotides of flanking sequence on either side, to allow for ambiguous origin of these bases. We identify as a possible D-gene match every maximal non-overlapping alignment to this segment of the three D-gene alleles. These D-gene matches are scored by their length and the top 200 are selected as possible D-gene alignments.

Alignment files are available online[8]: the files are in Matlab format and record the outcome of the above alignment strategy for a subset of our data. Inspection of the alignment data for individual sequences should provide instructive illustrations of the above-described procedure. The various thresholds and parameters used in the procedure are found in the files as well. The full set of alignment files used in our analysis can be generated using routines provided in our online software repository.

We note that one could generate a unique assignment of sequence features to a given read by selecting from the alignment ensembles just described the V, D, and J assignments with the highest score (i.e. having the longest effective alignment with the read). We will call the occurrence distribution of gene assignments, insertions, and deletions produced in this way as the 'deterministic ' estimate of the sequence feature probability distribution. It corresponds to standard practice in the literature for inferring feature statistics from sequence data, and will be used as a benchmark for comparison and contrast with our more accurate probabilistically inferred distribution.

[8] `www.princeton.edu/~ccallan/TCRPaper/Alignments`

2.6.4 Software

The algorithms we have developed to execute these two steps are described in greater detail in the following two subsections. Software to implement these procedures was written in Matlab using the Parallel Computing toolbox and run on a Linux cluster. Compiling key routines into C++ using Matlab Coder greatly improved processing speed, allowing model inference on an individual data set to be completed in about 20 hours running on 8 processors. Our Matlab code, along with summary instructions on how to run it, is available online[9].

2.6.5 Sequencing error rate

The sequence mismatch rate in our model reflects both uncorrected sequencing error as well as unknown allelic variation. Our model assumes that this mismatch rate R is independent of position along the sequence read. As is well-known, accuracy of the sequencing procedure becomes worse at the end of the sequence read (the $5'$, or V-gene, end of our CDR3 sequence) so, in assaying error rates, we ignore the last 15 nucleotides (at the $5'$end) for the 101 bp reads, where we can afford to do this. Our alignment procedure also disallows mismatches in the J- and D-gene alignment because of the shortness of these segments and the expected low error rate at this end (more accurately, the beginning) of the sequence read. In assessing position dependence of sequence error rates, therefore, we only need concern ourselves with mismatches to V gene assignments. Summing all such mismatches for the three individuals for which we have 101 bp reads, and plotting them against read position, we obtain the results plotted in Fig. 2.14. We find that R converges in the mean to a value of order 3×10^{-4} per base pair, two orders of magnitude smaller than the raw instrumental sequencing error rate. There are, however, a few sharp peaks at specific positions along the read; since they appear at the same position

[9] `www.princeton.edu/~ccallan/TCRPaper/Software`

for different individuals, they presumably reflect some anomaly in the functioning of the sequencing machine. This shortcoming of the error rate model does not greatly influence the results of the inference because the overall error rate is rather low.

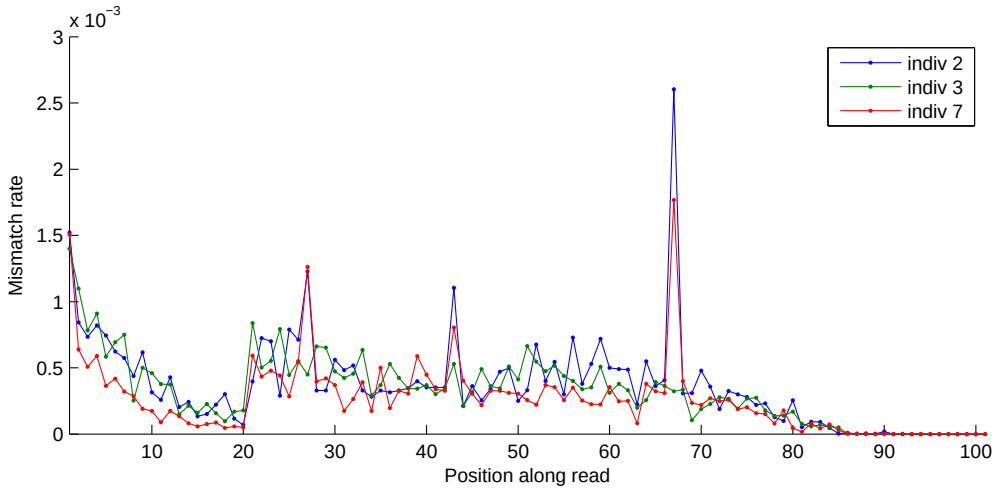

Figure 2.14: Position-dependent error profile for the three individuals with read length 101 base pairs. The sequencing read proceeds from the right (101 to 1) where the J gene sequencing primer binds. The spikes in the error rate at specific positions (67, 43 and 27) are true sequencing error spikes and not the result of unknown allelic variants. Positions 1-15 show the characteristic increase in error rate with read length. The overall decreased error rate in positions 10-20 reflect our requirement of a minimum alignment length of 20 nucleotides to a V gene with an upper bound on the allowed errors in the alignment. Since we do not allow any errors in the J and D genes, the error rate is zero in this region.

2.6.6 Spurious shared sequences between repertoires

Of the 9 individuals, we find three specific pairs of individuals – (2,3), (2,7) and (5,6) – who have an unusually large number of sequences in common, in both the naive and memory compartments. While all other pairs of individuals have between 0 and 4 sequences in common, these three pairs have 15 to 90 shared sequences. Additionally, many of these shared sequences occur in both the naive and memory compartments of the individuals. We suspect that these anomalies are the result of inter-sample

contamination.

Hence, for our analysis of the distribution of shared sequences between individuals, we discard from consideration the four pairs of individuals (2,3), (2,7), (3,7) and (5,6). This leaves 32 pairs of individuals for our analysis. We also discard three specific additional sequences that occur in the naive and memory compartments of one individual and also in another individual.

2.6.7 Sequence dependence of nucleotide deletion probabilities

Since the sequence at the $3'$ end of the V gene varies between genes, we fit a simple model to the gene dependent deletions profiles to explain the variation in these distributions. The precise mechanism of the generation of P-nucleotides and their relationship to deletions is unclear. Hence, we take only the probabilities of deletions greater than or equal to two nucleotides and consider the nucleotide sequence context (four bases $3'$ and two bases $5'$ of the deletion position) as a predictor of the deletion probability. We use a function of the form

$$P(n\,\text{deletions}|\sigma\,\&\,n \geq 2) = \frac{\exp\left(\sum_{k=1}^{6}\epsilon(k,\sigma(n-4+k)\right)}{Z(\sigma)} \tag{2.8}$$

$$Z(\sigma) = \sum_{n=2}^{12}\exp\left(\sum_{k=1}^{6}\epsilon(k,\sigma(n-4+k)\right) \tag{2.9}$$

where ϵ is a 6×4 matrix containing the contribution of each possible nucleotide at each of the positions, analogous to a (log) Position Weight Matrix (PWM). We do a least squares fit to determine the elements of ϵ. In Fig. 2.15, we show ϵ fit to the V deletions. There is a strong preference for T and A, especially in the 2 nucleotides just $5'$ of the position of deletion. Since there are only 13 J-genes, there is less sequence variation among them that we can utilize.

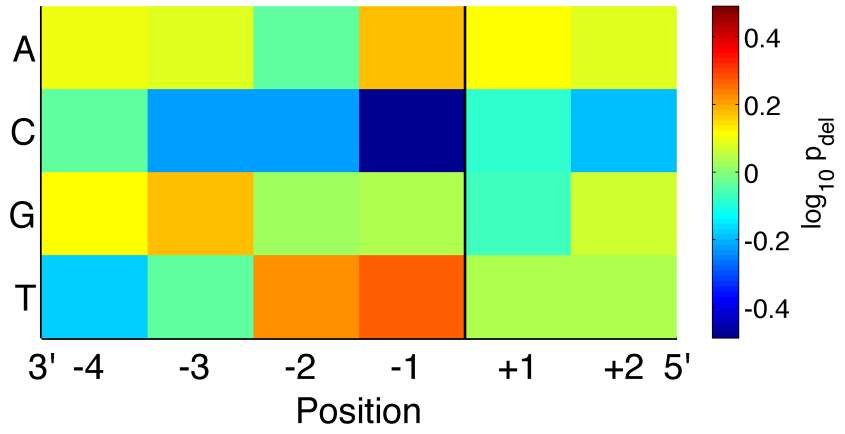

Figure 2.15: Position weight matrix for sequence dependence of nucleotide deletion position. The figure shows $\epsilon/\log(10)$ (see Appendix section 2.6.7) fit to the V gene specific deletions profiles, using four nucleotides $3'$ and two nucleotides $5'$ of the deletion position (black vertical line). The $3'$ nucleotides are the most informative about deletion probability and show a preference for T and A. The sequence logo corresponding to this position weight matrix is shown in the Fig. 2.9B.

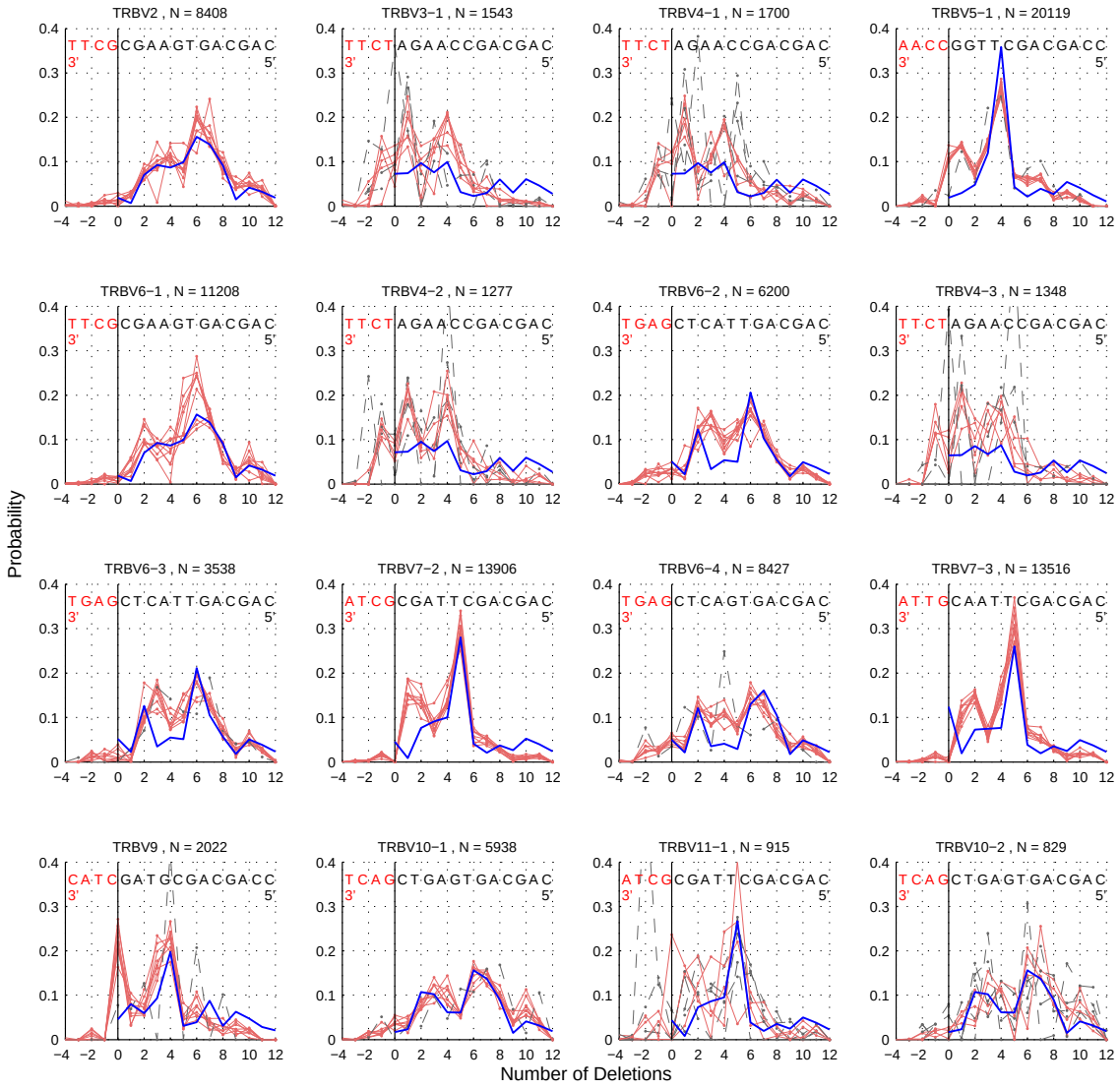

Figure 2.16: Deletion profiles for all the V-genes (1 of 3). The title for each panel lists the gene name and total number of counts, across all the individuals studied, of the particular gene in question. Individuals with fewer than 100 counts for a specific gene are plotted in gray dashed lines. The blue lines show the predictions of the position weight matrix based model fit to these curves.

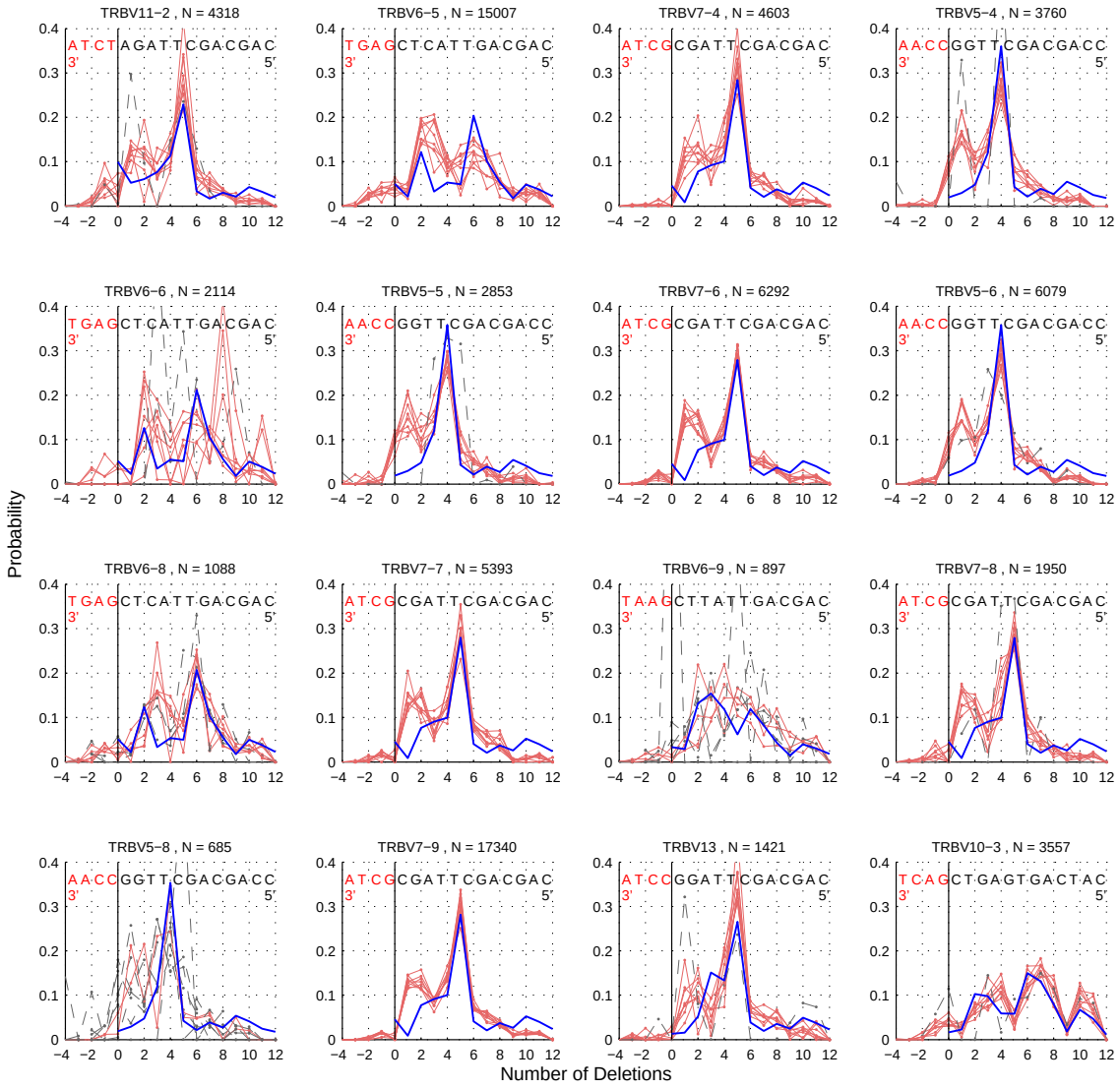

Figure 2.17: Deletion profiles for all the V-genes (2 of 3). The title for each panel lists the gene name and total number of counts, across all the individuals studied, of the particular gene in question. Individuals with fewer than 100 counts for a specific gene are plotted in gray dashed lines. The blue lines show the predictions of the position weight matrix based model fit to these curves.

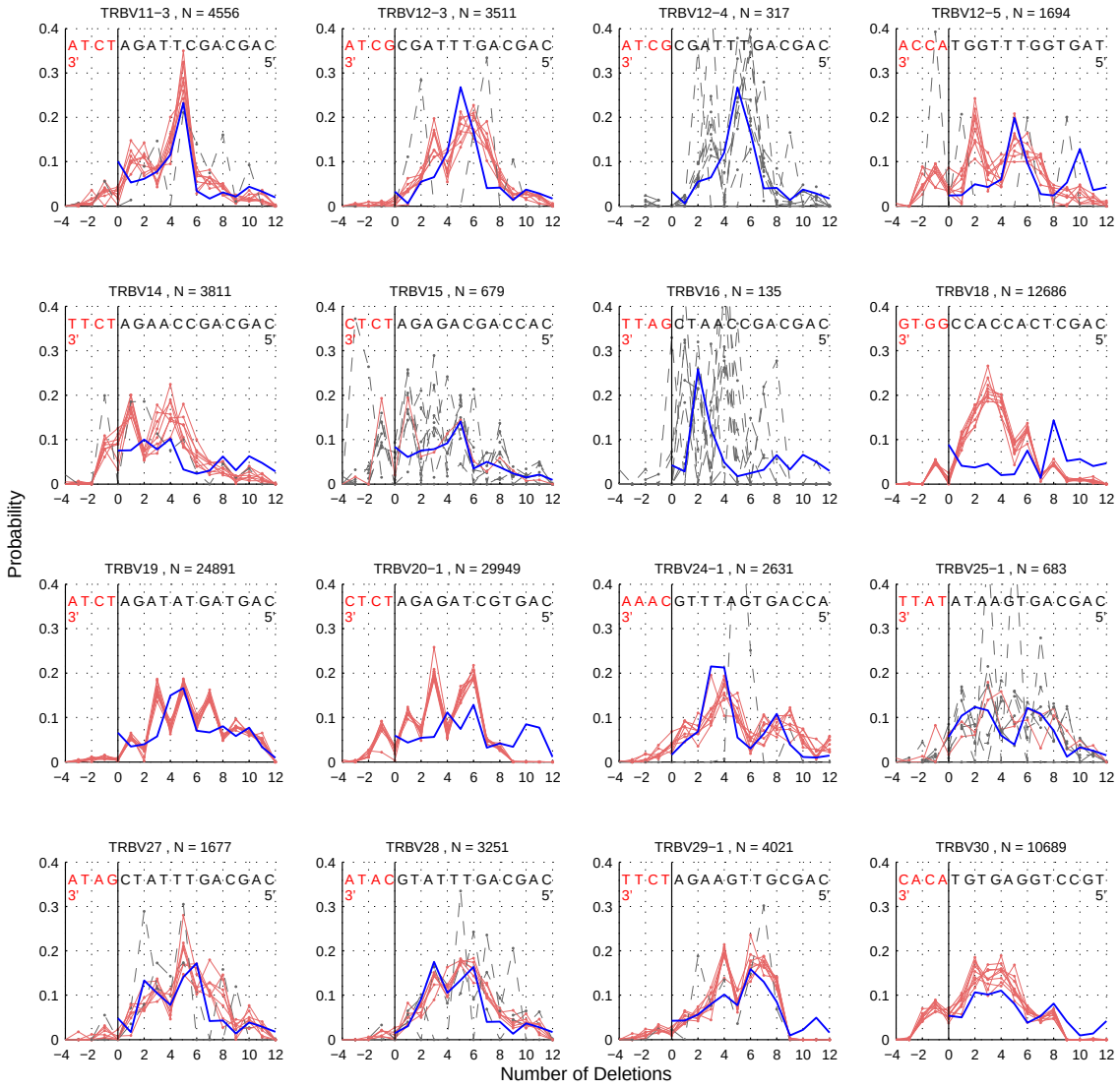

Figure 2.18: Deletion profiles for all the V-genes (3 of 3). The title for each panel lists the gene name and total number of counts, across all the individuals studied, of the particular gene in question. Individuals with fewer than 100 counts for a specific gene are plotted in gray dashed lines. The blue lines show the predictions of the position weight matrix based model fit to these curves.

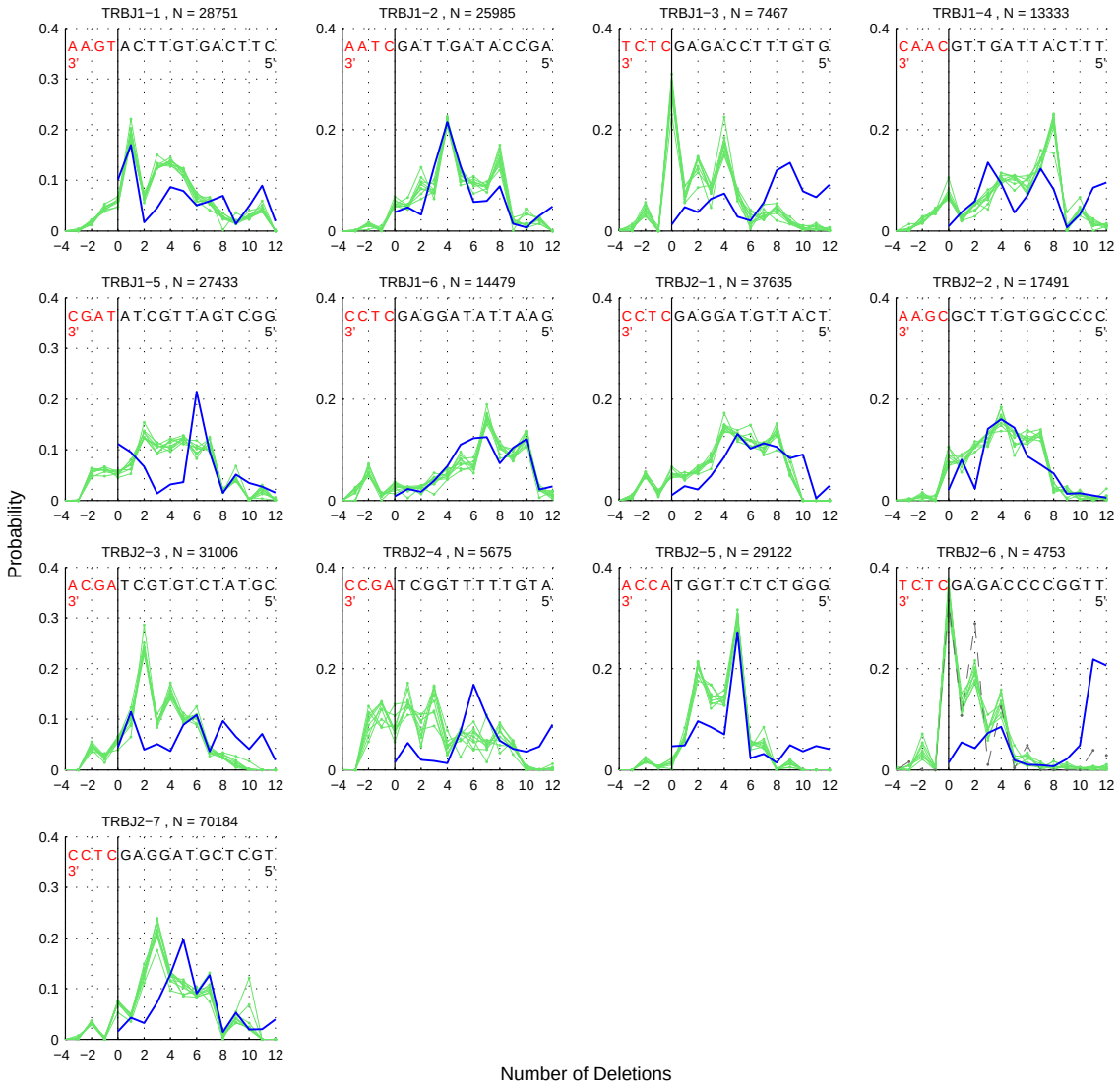

Figure 2.19: Deletion profiles for all the J-genes. The title for each panel lists the gene name and total number of counts, across all the individuals studied, of the particular gene in question. Individuals with fewer than 100 counts for a specific gene are plotted in gray dashed lines. The blue lines show the predictions of the position weight matrix based model fit to the V deletions curves, but evaluated on the J gene sequences.

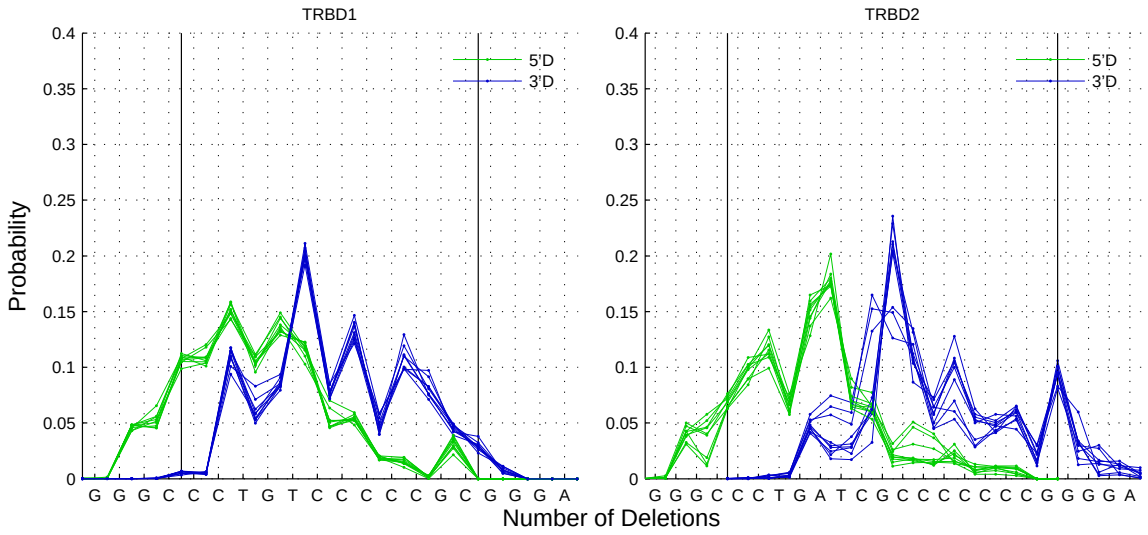

Figure 2.20: Marginal deletion probability distributions for the two D-genes. Deletions at the 5′end (3′end) of the D gene are shown in green (blue). The x-axis displays the gene sequence from the 5′end to the 3′end.

Chapter 3

Regulatory Sequences

3.1 Introduction

This chapter is an edited version of two publications [41,42]. The work described was done in collaboration with Justin Kinney, Curtis Callan and the co-authors listed in [42].

Across organisms and within organisms, large segments of DNA are homologous to each other. The biochemical variation in the proteins coded for by these genes contributes to the observed differences between species but cannot possibly fully account for the observed phenotypic diversity. Rather, the primary differences between species are in the relative expression levels of the various genes and the precise dynamics of these levels over time. This is demonstrated by the highly conserved proteins involved in embryonic development where the differences in morphology result from mutations that alter the schedule, location and quantity of their expression [40].

The regulation of gene expression must also necessarily be coded in the genome to be inherited. Multiple regulatory mechanisms have been discovered and act at various levels of the molecular process of gene expression. At the translational level, ribosome recruitment and elongation rates are influenced by RNA secondary structure which

is in turn specified by the sequence of the RNA itself. Additionally, noncoding RNAs can target transcripts from other genes (or other noncoding RNA molecules), greatly decreasing their expression levels.

At the transcriptional level, the recruitment of RNA polymerase to the transcription start site is highly influenced by the specific DNA sequences near the start site. Various proteins, called transcription factors (TFs), bind to DNA and interact with each other as well as with RNA polymerase, thus determining the rate of initiation of transcription. In eukaryotes, epigenetic information is another crucial determinant of transcription rates. Positioning of nucleosomes affects the accesibility of DNA to the transcriptional machinery. Regulatory sequences, called enhancers, can also be quite distant from their target genes in eukaryotes.

In this chapter, we focus on characterizing the molecular mechanisms of specific regulatory DNA sequences. Using data on the transcriptional activity of large libraries of mutant regulatory sequences, we infer models of their sequence-function relationships. In our work (section 3.2) on the *lac* promoter [41], we demonstrate a general experimental design and analysis strategy that can be applied to infer the functional binding sites on a regulatory sequence and quantitative models of their interactions. The experiments for this work were performed by Justin Kinney.

We then use a similar strategy to investigate two mammalian enhancers (section 3.3), which show greater complexity than the bacterial promoter and we use our model to engineer sequences of these enhancers to optimize their function [42]. These experiments were performed by collaborators at the Broad Institute.

3.2 Prokaryotic regulatory sequences

In prokaryotes, transcriptional regulation is implemented primarily by the DNA sequences that are near the $5'$ end of the target gene. Immediately upstream of the gene is a 'promoter' that contains a binding site for RNA polymerase and potentially other DNA binding proteins. These proteins are generally called transcription factors (TFs) and influence the transcription rate through their interactions with RNA polymerase and with each other [43]. A TF that has a favorable interaction with RNA polymerase increases the local concentration of the polymerase by binding to DNA near the promoter. TFs can also bind to DNA and sterically exclude RNA polymerase or other recruiters of the polymerase, thus repressing transcription.

3.2.1 The *lac* promoter

A classic example of transcriptional regulation in prokaryotes is that of the *lac* operon in E. coli, involved in lactose metabolism [44]. Figure 3.1 shows a cartoon of the *lac* operon. The operon contains three structural genes responsible for lactose transport and digestion. Expression of these genes is controlled by upstream regulatory sequences. The promoter region contains binding sites for σ^{70}-dependent RNA Polymerase (RNAP) and the transcription factor cAMP Receptor Protein (CRP). CRP acts as a recruiter of RNAP through its interaction with the α-subunit of RNAP, thus increasing the transcriptional activity. The active form of CRP has one or two molecules of cyclic AMP (cAMP) bound to it, which then enables the binding of CRP to DNA.

3.2.2 Thermodynamic models of promoter action

A widely used model of promoter mechanism is the so-called thermodynamic model [45–47] . The basic assumption of this model is that the rates of binding and unbinding

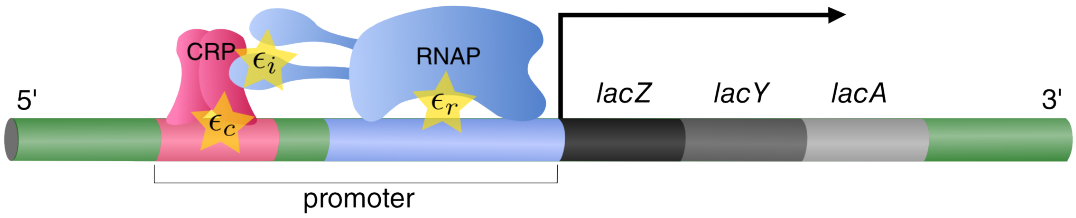

Figure 3.1: The *lac* operon in *E. coli*. The transcription of the genes *lacA*, *lacY* and *lacZ* is regulated by the promoter region which contains binding sites for CRP and RNAP. CRP interacts with the α-subunit of RNAP.

of the DNA binding proteins and RNA polymerase (typically > 1 sec^{-1}) are much higher than the rate of initiation of transcription by the DNA-bound transcriptional machinery (typically ~ 1 min^{-1}) [48]. This separation of time scales implies that the rate of transcript initiation must be proportional to the mean occupancy of RNA polymerase at the promoter.

The occupancy of DNA bound proteins (including the polymerase) are determined by their sequence dependent DNA-binding energies as well as the interactions between the proteins. Using this framework, for the *lac* promoter in Figure 3.1, we can model the transcription rate τ using the thermodynamic occupancy of RNAP at its binding site (see Eqn. 3.13).

Such thermodynamic models have been used to explain the dependence of transcriptional activity on the concentrations of transcription factors. In Kuhlman et al., [49], the up-regulation of transcription by the protein CRP was quantitatively explained by the model. To do this, Kuhlman et al. measured transcriptional activity resulting from different in vivo concentrations of active CRP and showed that the resulting functional form of this activity was consistent with such a model. In general, we do not have control over the in vivo concentrations of various transcription factors and often do not even know the identities of the TFs involved at a specific lo-

cus. Our work demonstrates a strategy, that we call 'Sort-Seq' to infer the biophysical mechanism of a specific regulatory sequence from measurements of the transcriptional activity of large libraries of mutant sequences.

3.2.3 Sort-Seq Experimental design

We start with the wild type promoter sequence and synthesize mutants that differ at each nucleotide position from the wild type nucleotide at a set mutation rate. Figure 3.2 illustrates the basic experimental protocol. Each mutant sequence is cloned into a reporter construct where the expression of a fluorescent reporter gene is driven by the regulatory sequence. These constructs consist of very low copy number plasmids that contain the *lac* promoter and a green fluorescent protein gene. This large library of plasmids is introduced into cells and expression of the reporter is induced.

The cells then display different levels of fluorescence due to differing transcriptional activities. They are then sorted by a fluorescence-activated cell sorting machine into 'batches' based on the intensity of fluorescence. A sample of cells from each batch are then taken and the mutant sequences in them are sequenced using ultra-high-throughput DNA sequencing. The final data set thus consists of a list of mutant sequences along with their batch numbers reflecting the level of transcriptional activity.

3.2.4 Data

Six Sort-Seq experiments were done. Table 3.1 lists the data sets that were generated. Figure 3.3 shows the mutation rates for three of the libraries. The data sets named full-wt, full-500, full-150 and full-0 have an average mutation rate of 12% per nucleotide over the whole 75 base pair sequence. Data sets crp-wt and rnap-wt have mutations only within the binding regions of CRP and RNAP respectively. Figure 3.4 shows the fluorescence distributions for the libraries before sorting along with the

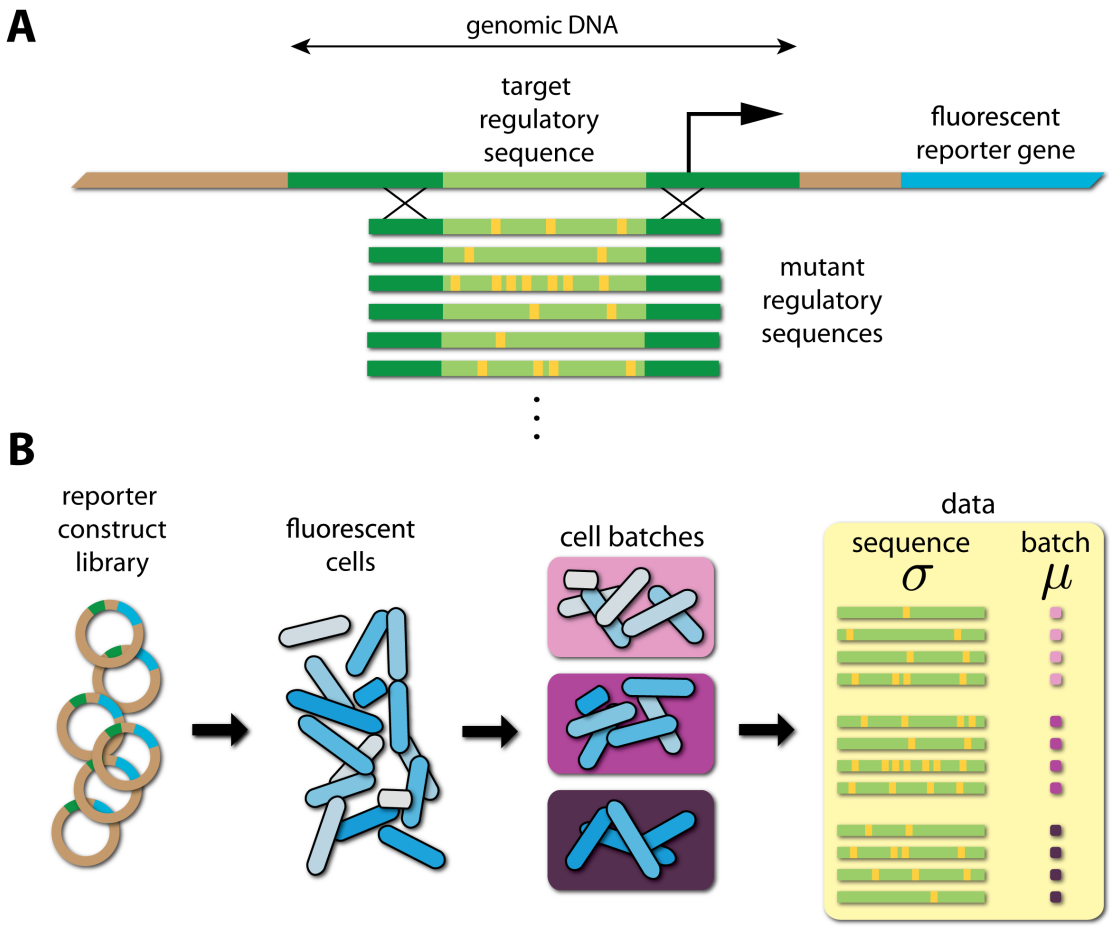

Figure 3.2: Sort-Seq Experimental design. A) The reporter construct with the mutant regulatory sequence inserted upstream of a fluorescent reporter gene. B) The reporter constructs are introduced into cells and their expression is induced. The fluorescent cells are then sorted into batches based on the intensity of fluorescence. Samples of cells from each batch are taken and the sequence in the reporter construct is determined by high-throughput sequencing. The final data set consists of a library of mutant promoter sequences with associated batch numbers. Figure from [41].

boundaries of the batches.

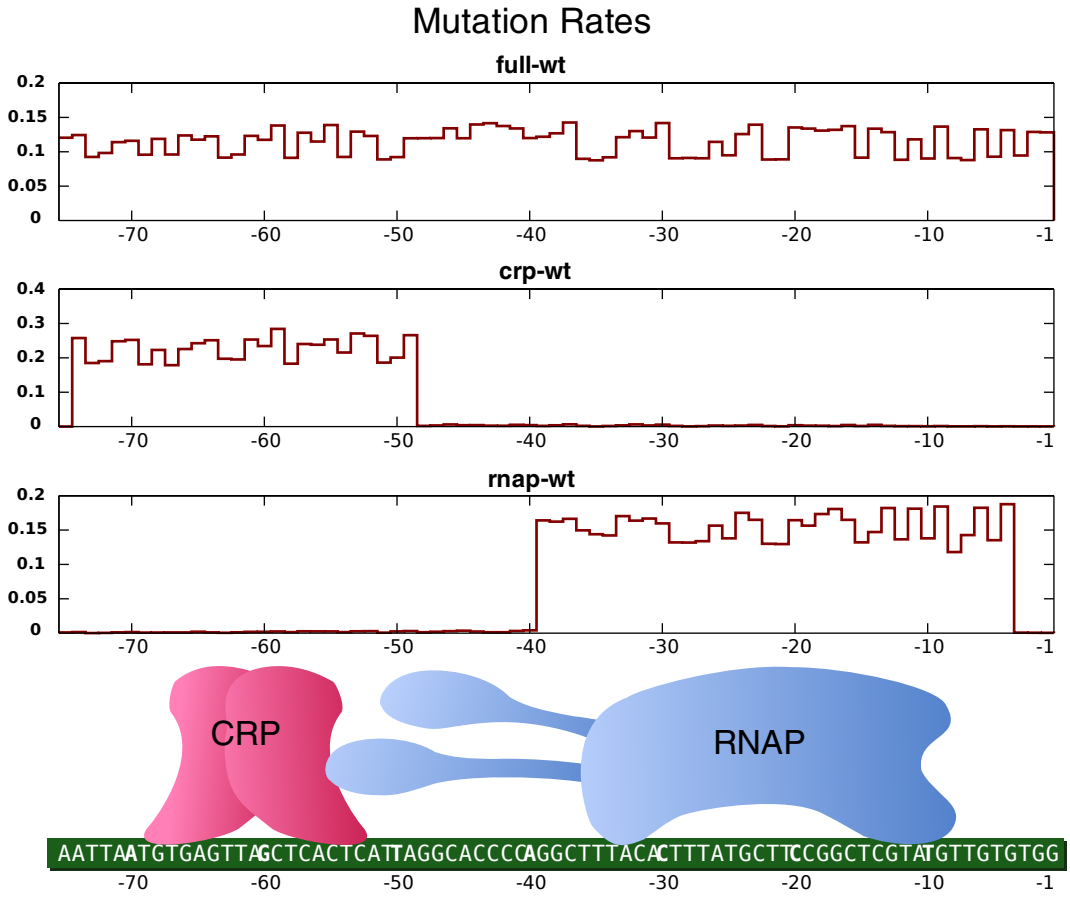

Figure 3.3: Mutation profiles of three Sort-Seq libraries. The average substitution rates are 12% for full-wt, 24% for crp-wt and 15% for rnap-wt. Only the CRP binding region [-74:-49] and the RNAP binding region [-39:-4] were subject to mutation in crp-wt and rnap-wt libraries respectively.

The wild-type strain (MG1655) of *E. coli* was used for the data sets full-wt, crp-wt and rnap-wt while a mutant strain (TK310) [49] that cannot produce or degrade cAMP was used for full-500, full-150 and full-0. The internal concentration of cAMP in this strain is determined by the cAMP concentration in the growth medium. This concentration was varied among the three data sets, allowing us to probe its effect on promoter activity.

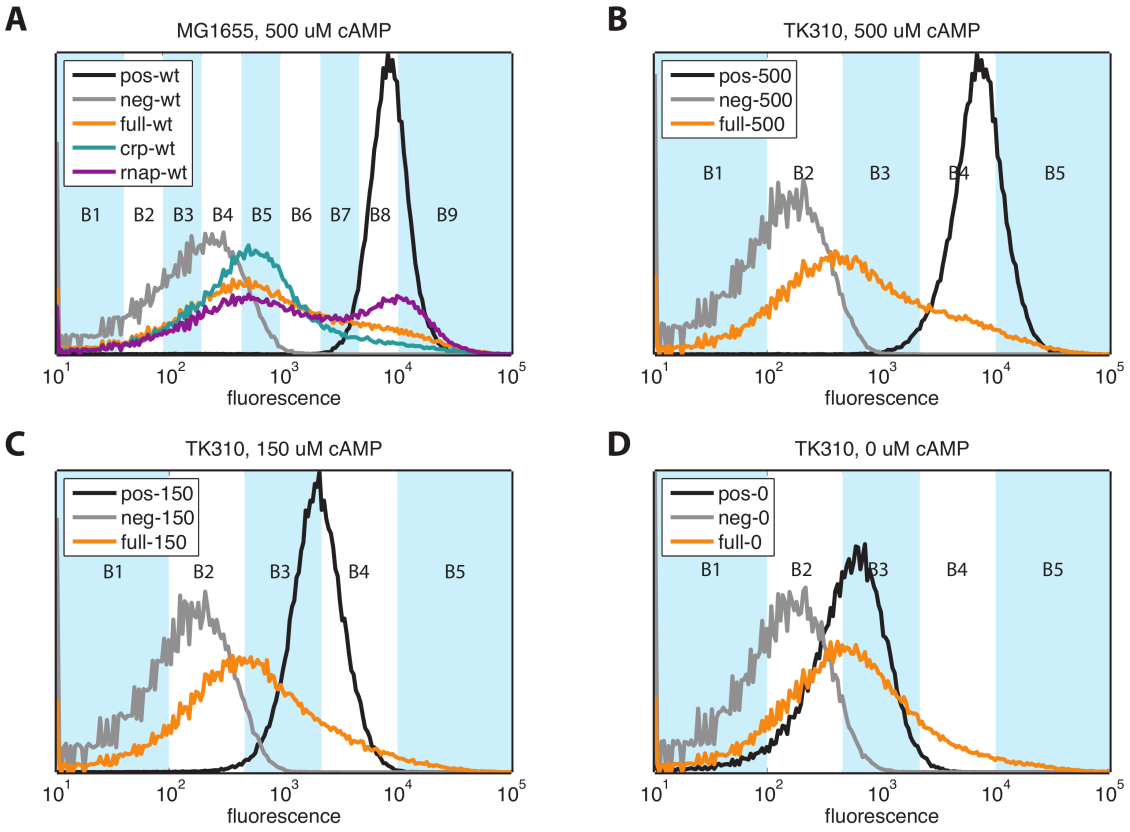

Figure 3.4: Fluorescence distributions before sorting. The 'pos' libraries (black lines) and the 'neg' libraries (grey lines) are controls. The promoter sequence is the wild-type sequence without mutations for 'pos' while the 'neg' libraries have the promoter sequence completely deleted. The boundaries of the batches are shown. For full-wt, crp-wt and rnap-wt experiments, the last batch (B10) was sampled from the entire distribution. (Figure from [41])

Data Set	Mut. Region	Mut. Rate	Strain	[cAMP]	Batches	Sequences
full-wt	[-75:-1]	12%	MG1655	500 μM	10	51,835
crp-wt	[-74:-49]	24%	MG1655	500 μM	10	46,986
rnap-wt	[-39:-4]	15%	MG1655	500 μM	10	45,461
full-500	[-75:-1]	12%	TK310	500 μM	5	23,431
full-150	[-75:-1]	12%	TK310	150 μM	5	24,334
full-0	[-75:-1]	12%	TK310	0 μM	5	28,544

Table 3.1: Sort-Seq data sets. The full-wt, crp-wt and rnap-wt libraries used the wild-type strain of *E. coli* while the other three used a mutant strain that cannot regulate cAMP concentration. The full libraries have the entire promoter region subject to substitution while crp-wt and rnap-wt have only the CRP binding region and RNAP binding region mutated. The concentration of cAMP in the growth medium of the full-500, full-150 and full-0 libraries was varied. The last column lists the number of sequences in the final data set after quality filters.

3.2.5 Statistical inference

Predictive information

Given a data set $\{\mu_i, \sigma_i\}_{i=1}^{N}$ of measured activity μ_i for each sequence σ_i, we build a model, with parameters θ, for an underlying biological quantity x that is informative about the measurements. To guide the model construction and fitting process, we define the mutual information [50], [51] between the measurements μ and the model predictions x as the 'predictive information', $I(x;\mu)$, of the model. Explicitly,

$$I(x;\mu) = \sum_{x,\mu} f(x,\mu) \log \frac{f(x,\mu)}{f(x)f(\mu)} \tag{3.1}$$

where $f(x,\mu)$ is the joint probability distribution of x and μ. This information is a measure of correlation that is insensitive to the actual functional relationship between the variables involved. Models that account better for the data will have a higher predictive information. Additionally, we can also set an absolute upper bound for the value of the predictive information, as explained below.

We can view the processes that result in a transcription rate for each sequence as steps in a signal compression chain. When a transcription factor binds to DNA, it

compresses all of the information in the sequence of nucleotides into a single quantity - the physical binding energy. Further, the binding energies of the transcription factors and the RNA Polymerase involved are then combined to produce a single thermodynamic transcription rate. Each of these steps passes on the information contained in the DNA sequence about the final measurement.

If the biological quantity being probed - in our case the transcription rate - is $x*$, a model of the noise in the experiment is provided by the conditional probability distribution $E(\mu|x*)$ of making the measurement μ while the underlying quantity has value $x*$. For each sequence σ, we have the following Markov chain.

$$x \xleftarrow[\text{model}]{x(\sigma,\theta)} \sigma \xrightarrow[\text{actual value}]{x*(\sigma)} x* \xrightarrow[\text{noise}]{E(\mu|x*)} \mu \tag{3.2}$$

Since mutual information obeys the data processing inequality [51], it follows that

$$I(x;\mu) \leq I(\sigma;\mu). \tag{3.3}$$

Estimating $I(\sigma;\mu)$ provides an upper bound for the predictive information. We therefore try to find models that maximize predictive information.

Sampling distribution

To infer the model parameters with uncertainties, one might sample parameter values from a likelihood distribution. Assuming a specific model of the noise in the experiment, $E(\mu|x)$, we can write the likelihood of model θ as

$$p(\theta|\{\mu_i\}) \sim p(\{\mu_i\}|\theta) = \prod_{\mu,x} E(\mu|x)^{Nf(x,\mu)} \tag{3.4}$$

$$= e^{N[I(x;\mu)-H(\mu)-D(f||E)]} \tag{3.5}$$

where $Nf(x,\mu)$ is the number of sequences with measured activity μ and predicted activity x, $H(\mu)$ is the entropy of μ and $D(f||E)$ is the Kullback-Leibler divergence between the experimental distribution $f(\mu|x)$ and the assumed error model $E(\mu|x)$

$$D(f||E) = \sum_{\mu,x} f(x,\mu) \log \frac{f(\mu|x)}{E(\mu|x)}. \tag{3.6}$$

The first two terms in the exponent of Equation (3.5) do not depend on the error model $E(\mu|x)$ and the second does not depend on the physical model θ. Since we wish to avoid modeling the noise in the experiment, the approach we take here is to drop the third term and simply sample according to the distribution

$$p(\theta|\{\mu_i\}) \sim e^{NI(x;\mu)}. \tag{3.7}$$

Another approach one might take (see [52]) is to average the likelihood in Equation (3.5) over all possible error models $E(\mu|x)$ to obtain an Error-Model-Averaged (EMA) likelihood. It can be shown that for large N and for a large class of prior distributions on the space of error models, EMA likelihood and the mutual information based distribution in Equation (3.7) become identical [52].

In the case of fitting a single model to multiple data sets, the natural generalization of the distribution in Equation (3.7) is to use the product distribution

$$p(\theta|\{\mu_i\}) \sim e^{N_\alpha I_\alpha(x;\mu)}. \tag{3.8}$$

where α indexes the data sets.

MCMC algorithm

We use a Markov Chain Monte Carlo (MCMC) algorithm to sample model parameters from the desired distribution. The basic algorithm is as follows:

- Pick initial model parameters θ and evaluate the predictive information $I(x_\theta;\mu)$.
- Iterate the following steps:

  Perturb parameters $\theta \to \theta'$

  Evaluate $I(x_{\theta'};\mu)$

  Replace θ by θ' with probability $\min\left(1, e^{N[I(x_{\theta'};\mu)-I(x_\theta;\mu)]}\right)$

  Record θ

- When the sampling is stationary with respect to time, the resulting ensemble is from the distribution $p(\theta) \sim e^{NI(x_\theta;\mu)}$

To make the MCMC sampling more efficient, we also incorporated a parallel tempering (or replica exchange) algorithm [53]. In this method, multiple random walks in model space are performed simultaneously with each particle having a certain inverse temperature β_i, sampling from the distribution $e^{\beta_i NI(x_\theta;\mu)}$. Periodically, pairs of particles, say (i,j), are allowed to exchange positions in model space (or equivalently, exchange temperatures) with a probability

$$P(i \leftrightarrow j) = \min\left(1, e^{(\beta_i-\beta_j)N[I_i(x;\mu)-I_j(x;\mu)]}\right). \tag{3.9}$$

This exchange process favors the presence of low temperature particles at better models. Only samples from the lowest temperature particles, with $\beta = 1$, are allowed in our final ensembles. The other higher temperature particles with $\beta < 1$ serve to explore the parameter space with greater latitude and help the $\beta = 1$ particles escape from local maxima through the exchanges. The inverse temperatures $\{\beta_i\}$ were chosen heuristically to increase the flow of particles between the extreme temperatures.

Calculating $I(x;\mu)$

In computing the predictive information $I(x;\mu)$, we first bin the continuous variable x into $n_x = 100$ bins, to get the binned variable $\bar{x}$. This binning introduces ruggedness in the information landscape - numerous local maxima are present and make the MCMC sampling more difficult. To make the sampling easier, we convolve the distribution $f(\bar{x},\mu)$ with a gaussian of width $\sigma_x = 4$ bins (4% of n_x) to get the smoothed distribution $\hat{f}(\bar{x},\mu)$. This new joint distribution is then used to compute the predictive information as

$$I(x;\mu) = \sum_{\mu} \int dx \; f(x,\mu) \log \frac{f(x,\mu)}{f(x)f(\mu)} \tag{3.10}$$

$$= \sum_{\mu} \hat{f}(\mu) \sum_{\bar{x}} \hat{f}(\bar{x}|\mu) \log \frac{\hat{f}(\bar{x}|\mu)}{\sum_{\mu'} \hat{f}(\mu')\hat{f}(\bar{x}|\mu')} \tag{3.11}$$

The results obtained for the parameters did not vary significantly with the choice of σ_x or n_x.

3.2.6 Estimating $I(\sigma;\mu)$

To estimate the intrinsic information in the data $I(\sigma;\mu)$, we need to estimate the probability distribution $p(\mu|\sigma)$. We can then calculate

$$I(\sigma;\mu) = \left\langle \sum_{\mu} p(\mu|\sigma) \log \frac{p(\mu|\sigma}{p(\mu)} \right\rangle_{\sigma} . \tag{3.12}$$

If our data set contains enough independent measurements of μ for the same σ, we can get the conditional distribution $p(\mu|\sigma)$ from the data set directly. In our case, there are very few repeated sequences. Therefore auxiliary measurements were needed.

A small sample of sequences, about 10, from each batch was taken and the fluorescence distributions of a clonal population of each sequence was obtained. From

these distributions, the conditional distributions can be estimated after accounting for some selection biases. For instance, in the final data set there are roughly equal number of sequences in each batch even though before sorting, some batches have much larger fractions of the library. Thus the probability that a given sequence fluorescing within the boundaries of batch μ gets picked to be in the final data set varies from batch to batch. Additionally, the number of sequences in each batch of the final data set are only roughly equal. Both of these factors must be corrected for in going from the fluorescence distributions of the clonal populations of our small sample of sequences to an estimate of $p(\mu|\sigma)$.

3.2.7 Results

Information footprints

To reveal which specific sites in the promoter sequence are functionally important we can calculate an information footprint from the data. This is the mutual information between the nucleotide at a specific position and the measurement μ, i.e. $I(b_i;\mu)$. We find this for each position i to get a functional footprint. Figure 3.5 shows the information footprints for all six data sets with error bars.

The information values range over two orders of magnitude. Looking at the footprints of the full-500, full-150 and full-0 data sets, the effect of decreasing the cAMP concentration is immediately apparent. The footprint in the CRP binding region gets weaker and, for full-0, essentially goes to zero within error bars, with the exception of a few base pairs (-64, -57 and -52). It is also clear from the full-wt footprint that most positions in the sequence are informative about the activity of the promoter. Only 10 positions are consistent with no effect on expression. For less well-studied regulatory sequences, such footprints can be used to determine the exact locations of functional importance where transcription factors might bind.

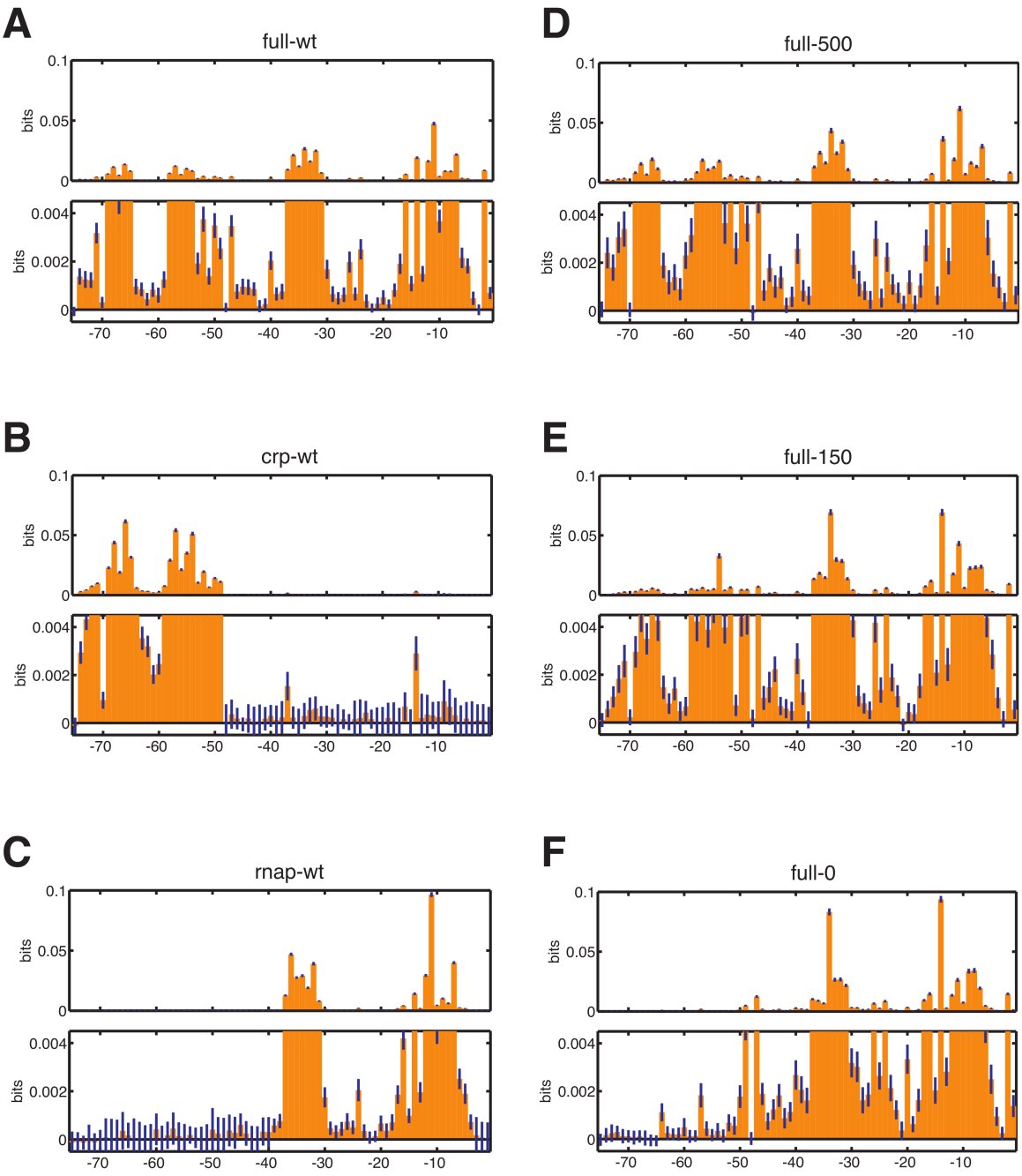

Figure 3.5: Information footprints for the six data sets. Error bars are shown in blue. The lower plot of each footprint is a 20X magnification of the upper plot. D, E and F reveal the effect of lowering the cAMP concentration. The CRP binding site becomes less functionally important and the footprint in this region goes to zero at all but three positions for full-0.

Thermodynamic model structure

As described in section 3.2.2, the transcription rate τ is assumed to be proportional to the thermodynamic occupancy of RNAP at its binding site given by

$$\tau = \tau_{max} \frac{C_r e^{-\epsilon_r/kT} + C_r C_c e^{-(\epsilon_r+\epsilon_c+\epsilon_i)/kT}}{1 + C_c e^{-\epsilon_c/kT} + C_r e^{-\epsilon_r/kT} + C_r C_c e^{-(\epsilon_r+\epsilon_c+\epsilon_i)/kT}}. \tag{3.13}$$

Here ϵ_r and ϵ_c are the DNA binding energies of RNAP and CRP respectively, while ϵ_i is the (sequence independent) interaction energy between the two. C_c and C_r are the concentrations of CRP and RNAP respectively. The sequence dependent DNA binding energy of proteins can be modeled using 'energy matrices'. Each base pair contributes independently to the total binding energy, an amount that depends on the nucleotide present. This simple model ignores non-linear contributions like the DNA bending energy, but is quite successful for many transcription factors and widely used [54, 55].

The concentration of RNAP cannot be determined since τ is a sequence-independently monotonic function of C_r and the predictive information is invariant under changes of such parameters. However C_c, the binding energy matrices and the interaction energy can be pinned down. Note also that the DNA binding energies can only be determined relative to the chemical potential, i.e. upto an overall sequence independent shift. This is because we can only probe the actual occupancy of the binding sites, not the concentration and the energy independently. If we set the binding energy of the wild-type sequence to zero, we can express the CRP concentration relative to the wild-type dissociation constant K_d^{wt}.

Thus the full thermodynamic model has two energy matrices, a CRP-RNAP interaction energy and the CRP concentration as parameters. To fit all the parameters in the thermodynamic model, we use the transcription rate τ as the model prediction x. However, it is also possible to fit just an energy matrix to either CRP or RNAP

by using the binding energy ϵ_c or ϵ_r as the model prediction x.

Final model parameters

For our final model, we combined the six data sets and used the model for the transcription rate τ with a single CRP-RNAP interaction energy ϵ_i but a different CRP concentration C_c for each data set. Single energy matrices were used across the data sets for CRP and RNAP. Figure 3.6 shows the results of this inference.

Heat maps of the two binding energy matrices in units of kcal/mol are shown along with the distributions of other parameters. The CRP-RNAP interaction energy was found to be $\epsilon_i = -2.8 \pm 0.1$ kcal/mol. This differs from the value of -3.37 ± 0.03 kcal/mol measured by Kuhlman, et al. in [49] by about 20%.

The ratios of the concentrations of CRP in the full-500, full-150 and full-0 data sets agree with the ratios of cAMP in the growth medium. This makes sense since CRP is activated by a single cAMP molecule, making the concentration of active CRP proportional to that of cAMP, and the internal cAMP concentration is proportional to the exogenous concentration [49]. The ratio of the inferred C_c for full-150 and full-500 was $10^{-5.6\pm0.2}$ while the ratio of cAMP concentrations was $\frac{150\mu M}{500\mu M} = 10^{-0.52}$. The full-0 experiment had a trace amount of cAMP ($\sim 50nM$) and this also agrees with the inferred ratio of C_c of full-0 to full-500 of $10^{-3.7\pm0.6}$, compared to the expected 10^{-3}.

Binding energy matrices

For each individual dataset, energy matrices were also obtained by maximizing $I(\epsilon_c; \mu)$ for CRP and $I(\epsilon_r; \mu)$ for RNAP where ϵ_c and ϵ_r are their respective DNA binding energies, with the exception of the full-0 dataset where the CRP concentration was very low. Since $I(\epsilon_c; \mu)$ and $I(\epsilon_r; \mu)$ are invariant under scaling of the binding energies by arbitrary factors, the matrices inferred this way can not be determined in physical

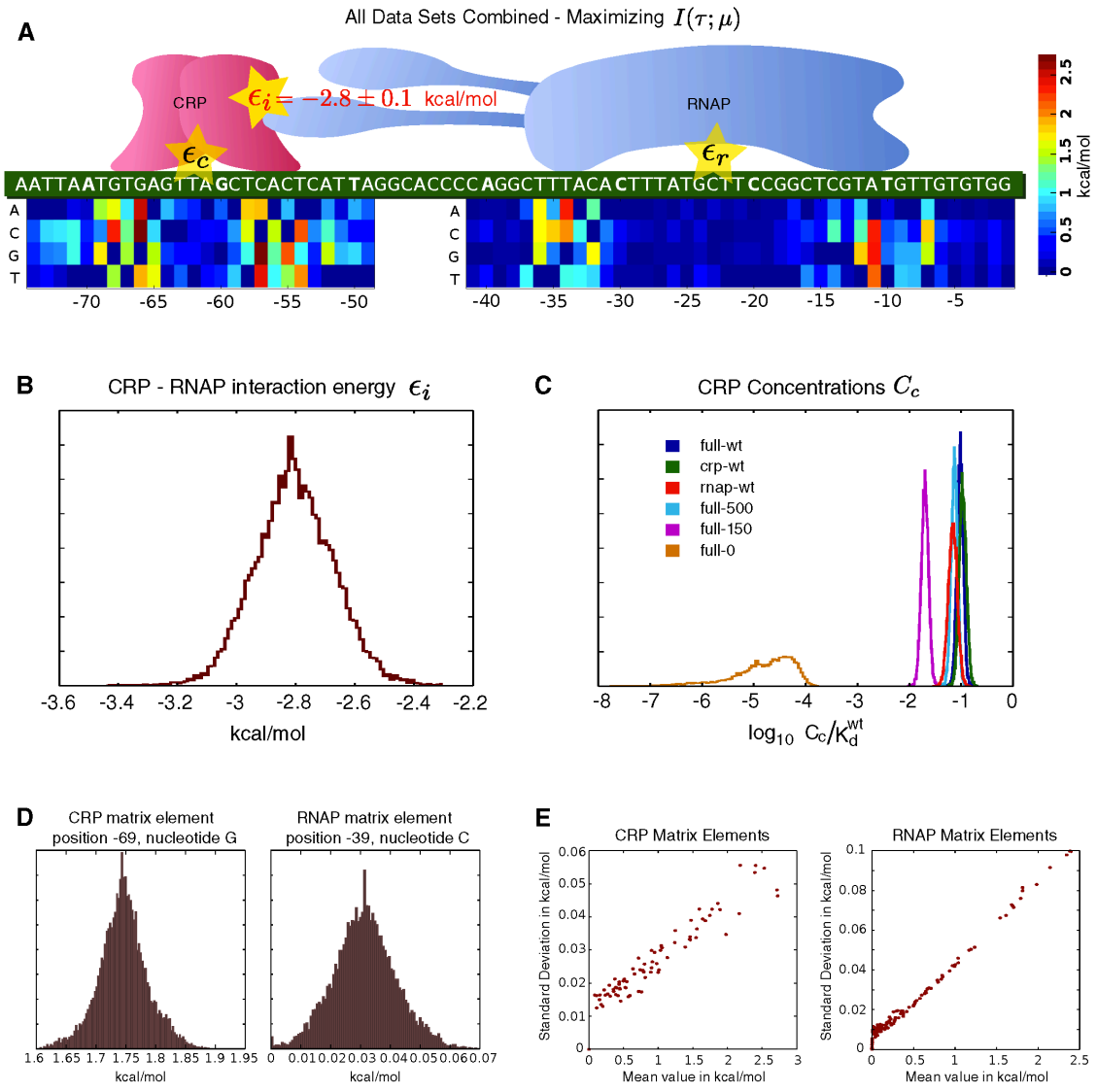

Figure 3.6: The final model inferred on the joint data set by maximizing $I(\tau;\mu)$. A) Energy matrices in units of kcal/mol for CRP (on the left) and RNAP (on the right). B) Distribution of the CRP-RNAP interaction energy ϵ_i. C) Distributions of the six CRP concentrations C_c relative to K_d^{wt}, the dissociation constant of the wild-type sequence. D) Distributions of sample energy matrix elements. E) Scatter plot of the standard deviation of energy matrix elements versus their mean values. Higher valued (unfavorable) elements have higher uncertainties as expected, since they matter less in determining specificity.

units. The overall scale is unknown. Figure 3.7B shows all of these results, normalized by the maximum element of each matrix. All the matrices from different data sets are seen to be extremely similar and are mostly consistent to within error bars.

Model completeness

Data set full-wt	
$\boldsymbol{I(\sigma;\mu)}$	$\mathbf{1.21 \pm 0.07}$ **bits**
$I(\epsilon_c, \epsilon_r; \mu)$	0.732 ± 0.006 bits
$I(\tau; \mu)$	0.732 ± 0.007 bits
$I(\epsilon_c + \epsilon_r; \mu)^\dagger$	0.647 ± 0.005 bits
Data set crp-wt	
$\boldsymbol{I(\sigma;\mu)}$	$\mathbf{0.88 \pm 0.09}$ **bits**
$I(\epsilon_c; \mu)$	0.727 ± 0.005 bits
Data set rnap-wt	
$\boldsymbol{I(\sigma;\mu)}$	$\mathbf{1.09 \pm 0.08}$ **bits**
$I(\epsilon_r; \mu)$	0.795 ± 0.005 bits

Table 3.2: Intrinsic information $I(\sigma;\mu)$ and maximum predictive information $I(x;\mu)$ for data sets full-wt, crp-wt and rnap-wt.
†This is the maximum value obtained by scaling ϵ_r with respect to ϵ_c.

Table 3.2 shows the maximum values of the predictive information obtained along with the intrinsic information for three data sets. Our model accounts for about 60% of the intrinsic information in the full-wt data set. However, the information contained in the pair of binding energies (ϵ_c, ϵ_r) is almost completely captured by the transcription rate. The information is substantially lowered if you replace the transcription rate by a linear sum of the energies. Thus this is a non-trivial confirmation of the validity of the thermodynamic model.

The remaining intrinsic information that is not accounted for could be due to several reasons. RNAP is known to have an alternative binding site [56] within the same promoter region. The α-subunit of RNAP also interacts with the DNA directly [57], increasing transcription. There could also be non-linearities in the DNA binding energies of CRP and RNAP from DNA bending, etc.

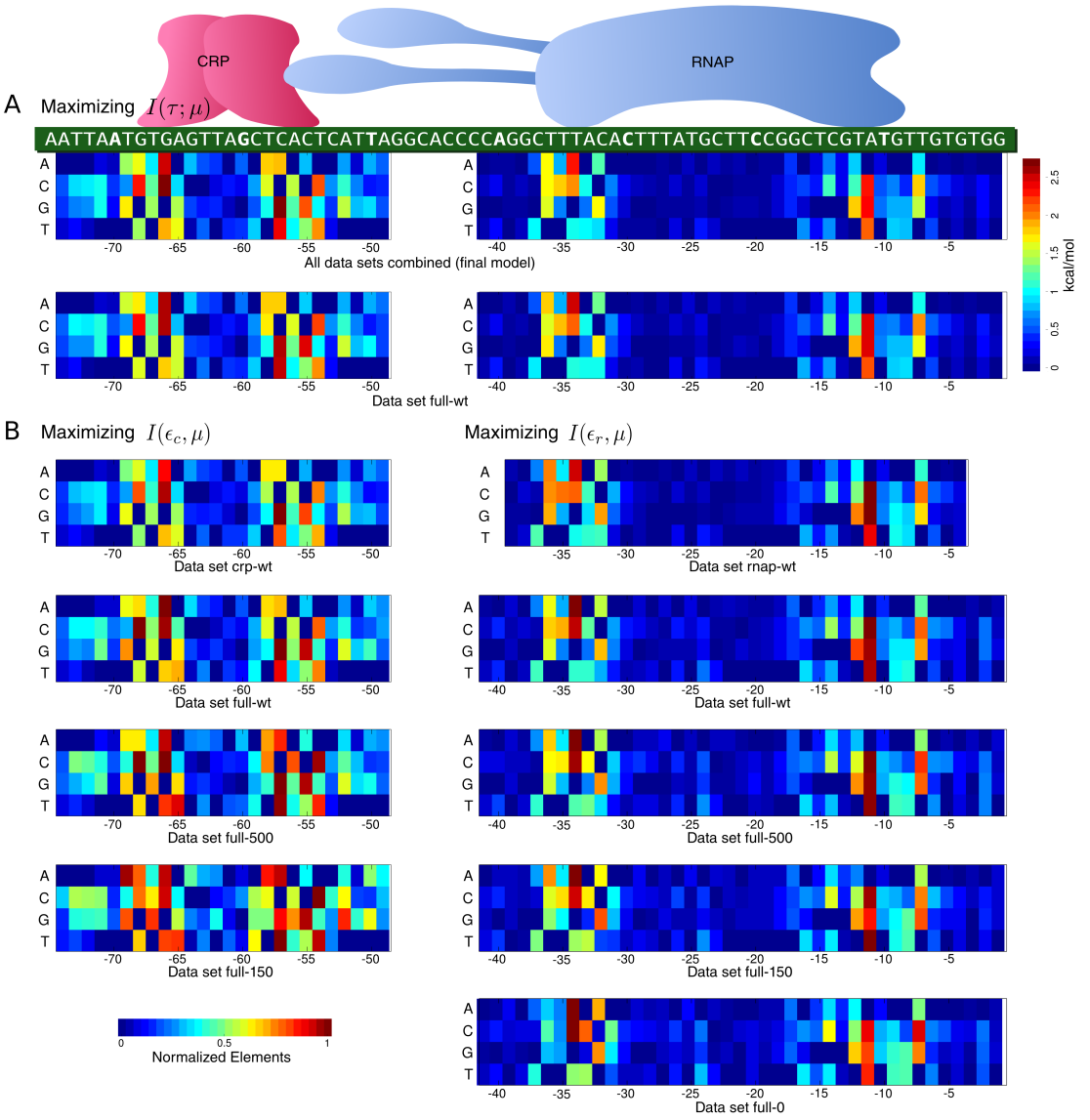

Figure 3.7: A) Energy matrices in units of kcal/mol for CRP (on the left) and RNAP (on the right), inferred by maximizing total $I(\tau;\mu)$ on the joint data set (first row) and just the full-wt dataset (second row). B) Energy matrices inferred by maximizing $I(\epsilon_c;\mu)$ for CRP and $I(\epsilon_r;\mu)$ for RNAP on each individual data set. These matrices are not in physical units and are determined only up to an unknown scale factor. Here they are all normalized so the highest element of each matrix is one. In both A and B, the lowest element of every column is constrained to be zero, since these elements contribute only to an overall shift in the energies. The rnap-wt energy matrix for RNAP is shorter because the mutagenized region for this dataset is [-39:-4].

3.2.8 Summary

The experimental method used in our work can be applied in a stereotyped manner to a wide variety of transcriptional regulatory sequences in a number of different organisms. No prior knowledge of how a sequence of interest functions is needed. All that is required is that the regulatory sequence of interest function on a reporter construct, and a large library of reporter constructs be introduced into cells so that each cell receives a single construct. The analysis techniques we describe here can then be used to identify all functional binding sites within the probed sequence region, characterize the sequence-dependent binding energy of the proteins that bind those sites, and then use this information to build biophysical models for how these proteins physically regulate transcription.

Using mutual information in model fitting provides great latitude in the design of such experiments: any experimental method that partitions mutant sequences according to function no matter how noisy the partitioning can be used in place of flow cytometry. For instance, sequencing the bands from a single gel shift experiment performed on a library of mutant binding sites should allow one to characterize the sequence-specificity of a DNA-binding protein with very high resolution. If multiple binding sites are included in each shifted oligo, such an experiment could also reveal the interaction energy between the proteins that bind these sites. Its likely that other commonly used low precision assays, such as SELEX, DNase I footprinting, and yeast 2 hybrid, can also be replaced with high-precision deep-sequencing-based alternatives. Indeed, some such methods have already been proposed [58,59]. The sequence analysis methods described here, however, allow quantitative models of arbitrary form to be inferred from such data with minimal assumptions about experimental details. As sequencing costs continue to fall and read lengths continue to increase, Sort-Seq and Sort-Seq-like assays should soon become feasible as standard laboratory techniques.

3.3 Eukaryotic regulatory sequences

In eukaryotes, transcriptional regulation is more complex. First, eukaryotic DNA is packaged by histones. The positioning and modifications of these nucleosomes greatly affects the accessibility of DNA for transcription [60–62]. Second, eukaryotic regulatory sequences can be tens of thousands of base pairs away from their target genes. These elements are called enhancers or silencers and typically strongly influence the expression of their target genes, presumably by looping and interacting with the promoter [4]. Additionally, so-called insulators are boundary elements that block the influence of enhancers on genes downstream of the insulator, limiting the domain of action of enhancers. Eukaryotic transcription factors are also typically have smaller DNA-binding footprints, between 4 and 12 base pairs, compared to the larger bacterial binding sites of 15-40 base pairs. In turn, typical eukaryotic regulatory sequences have many more protein binding sites, that often overlap with each other.

In this section we discuss the results of experiments and analysis similar to the previous section on two mammalian enhancers [42] : a synthetic cAMP-regulated enhancer (CRE), which is widely used as a cAMP sensor in cell signaling research and drug discovery [63], and the virus-inducible enhancer of the human interferon beta (IFNB) gene, which is one of the most comprehensively studied mammalian regulatory elements [64]. While the synthetic CRE is a 'billboard' enhancer composed of multiple nonoverlapping binding sites for the cAMP-responsive transcription factor CREB, the IFNB enhancer contains six different overlapping transcription factor binding sites. Thus their architectures and presumably the biophysical mechanisms of regulation are rather different.

3.3.1 Experimental design

Unlike the experiments in the previous section on bacterial promoters, here the transcriptional activity of mutant enhancers was measured by sequencing of mRNA transcripts. Briefly, our collaborators first synthesized tens of thousands of oligonucleotides that contain a library of regulatory elements, each coupled to a short tag. The oligonucleotides were used to generate a pool of plasmids, where each plasmid contains one of the regulatory elements, an optional invariant promoter, an arbitrary open reading frame (ORF) and an identifying sequence tag. The plasmids were co-transfected into cells, where active elements drive transcription of mRNAs containing the tags in their $3'$ untranslated regions. To estimate their relative activities, we sequenced and counted the tags in the reporter mRNAs and the plasmids pools, and then took the ratios of these counts.

3.3.2 Data

Our collaborators synthesized 142-mer oligonucleotide pools containing 87-nt CRE and IFNB enhancer variants, as well as 10-nt tags and various invariant sequences required for cloning. Two different mutagenesis strategies were tested. The first was 'single- hit' scanning where we assayed $\sim$ 1000 specific enhancer variants, including all possible single substitutions, multiple series of consecutive substitutions and small insertions at all positions. Each scanning variant was linked to 13 tags for a total of 13,000 distinct enhancer-tag combinations. This redundancy provides parallel measurements for each variant, which can be used to both quantify and reduce the impact of experimental noise, including tag-dependent bias.

The second was 'multi-hit' sampling where about 27,000 distinct enhancer variants were assayed, each linked to a single tag. These variants were constructed by introducing random nucleotide substitutions into the enhancers at a rate of 10% per position. Because the variants were designed in silico and then synthesized, they

provided a uniform mutational spectrum.

3.3.3 Results

Information footprints

As we did for the *lac* promoter in section 3.2, we calculate the information footprint from the multi-hit data for CRE and IFNB by estimating the mutual information between the nucleotides at each position and the corresponding tag ratios across the $\sim 27,000$ variants. These are shown in Figs. 3.8A and 3.9A.

We find that the 27 most informative positions in the induced CRE footprint are all located in or immediately flanking the four CREB sites (Fig. 3.8a). This clearly shows the primary importance of the CREB sites to the activity of the induced enhancer. The more symmetric footprint of dimeric CREB site 4 compared to site 1 is likely due to the extended palindromic flanks of the former (ATTGACGTCAAT vs. AGTGACGTCAGC). The information contents of CREB sites 2-4 were substantially lower in the uninduced state, which is consistent with cAMP- dependence. In contrast, the information contents of CREB site 1 and the cryptic binding sites near CREB sites 1 and 4 were higher in the uninduced footprint. This is again consistent with the most promoter-distal CREB site being less cAMP-dependent 14 and suggests that these sites may be particularly relevant for controlling the basal activity of the CRE.

The IFNB enhancer footprint from virus-infected cells shows, as expected, that its functionally relevant nucleotides are concentrated in the 44 nt core (Fig. 3.9A). Indeed, 35 of 46 positions that had significant information content are located in the core. Strikingly, the IFNB footprint from the uninduced state revealed only 8 informative positions, compared to 73 in the uninduced CRE footprint. This comparatively low information content likely reflects the near absence of transcription factor binding, which results in a very low basal activity (at least 5-fold lower than the uninduced CRE in luciferase assays).

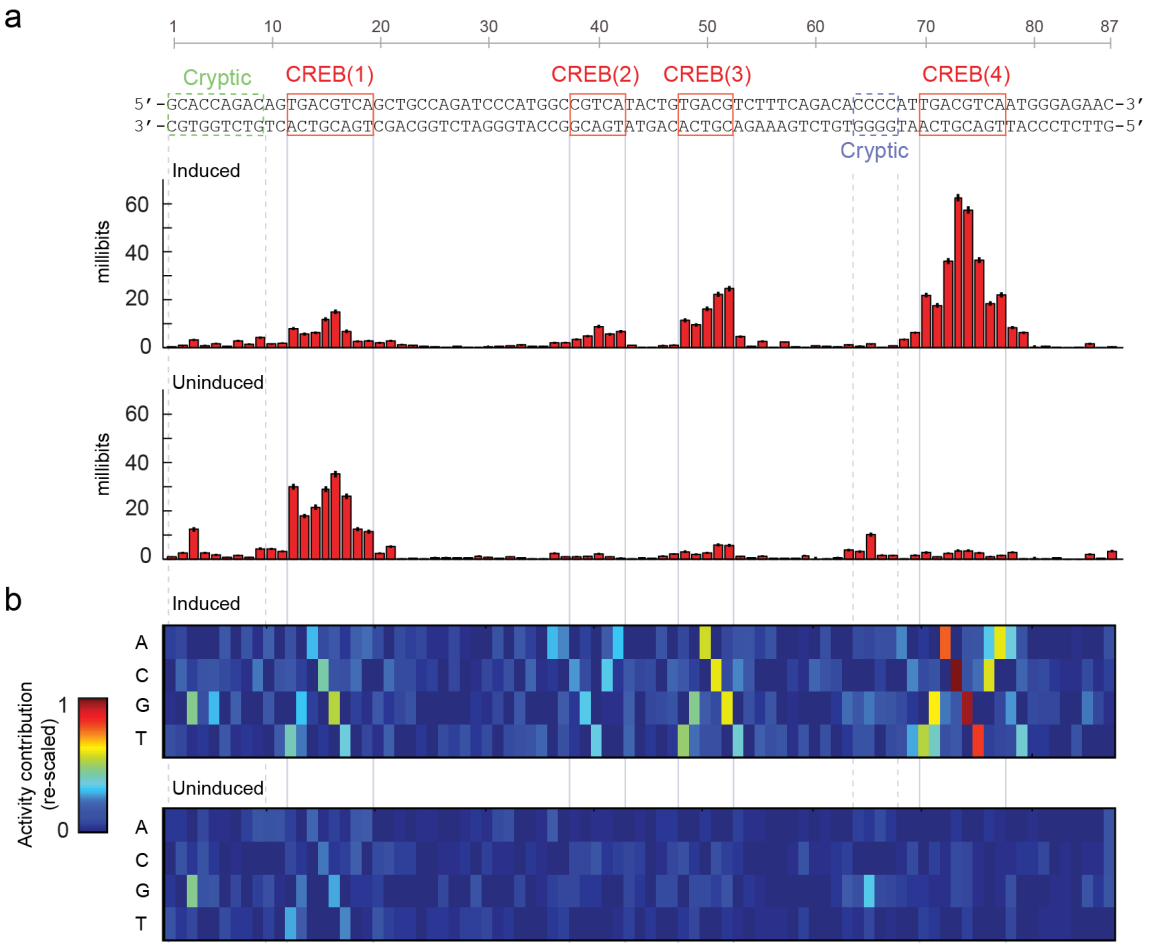

Figure 3.8: Multi-hit sampling mutagenesis of the cAMP-responsive enhancer. (a) Information footprints of the CRE in its induced (top) and uninduced (bottom) states. Red indicates significant information content at the corresponding position (permutation test, 5% FDR). Error bars show uncertainties inferred from subsampling. (b) Visual representations of linear models of the CRE in its induced (top) and uninduced (bottom) states. The color in each entry represents the estimated additive contribution of the corresponding nucleotide to the log-transformed activity of the enhancer. The matrices are rescaled such that the lowest entry in each column is zero and the highest entry anywhere is one. Both matrices are shown on the same scale. Figure from [42].

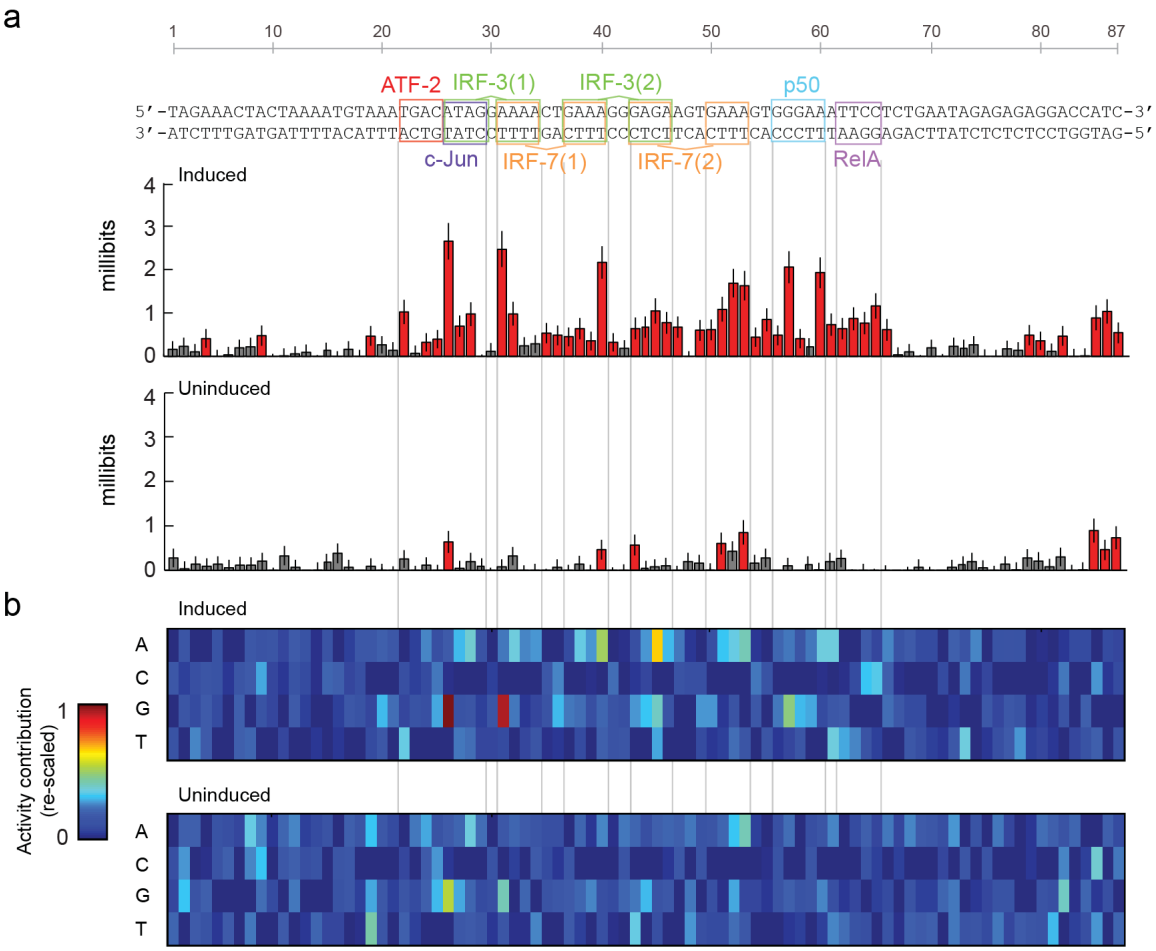

Figure 3.9: Multi-hit sampling mutagenesis of the virus-inducible IFNB enhancer. (a) Information footprints of the IFNB enhancer in its induced (top) and uninduced (bottom) states. Red indicates significant information content at the corresponding position (permutation test, 5% FDR). Error bars show uncertainties inferred from subsampling. (b) Visual representations of linear models of the IFNB enhancer in its induced (top) and uninduced (bottom) states. The color in each entry represents the estimated additive contribution of the corresponding nucleotide to the log-transformed activity of the enhancer. The matrices are rescaled such that the lowest entry in each column is zero and the highest entry anywhere is one. Both matrices are shown on the same scale. Figure from [42].

Models of sequence-dependent activity from multi-hit data

We attempted to develop quantitative models of the sequence dependent activity of the two enhancers, with the goal of predicting the activity of variants that were not assayed in our initial experiments. Unlike the physically motivated models that were successful in section 3.2, we were unable to find models that significantly outpeformed heuristic models of the sequence-activity relationship.

Specifically, we used linear regression to estimate the contribution of each possible nucleotide at each position to the log-transformed activity of each enhancer in their induced and uninduced states. The linear model is defined by parameters M_{bi} representing additive contributions of the different bases b at each enhancer position i to log transcriptional activity:

$$\log\ A(\sigma) = \sum_{b,i} M_{bi}\sigma_{bi} \tag{3.14}$$

where $A(\sigma)$ is the activity of sequence σ and σ_{bi} is 1 if σ has the nucleotide b at position i.

The model has 4 x 87 = 348 parameters, but because one of the four bases must be present at every position there are only 1+3 x 87=262 independent degrees of freedom. The primary virtue of these linear models is their simplicity, but it is not a priori obvious that such models can capture the complex response of multisite enhancers. Nonetheless, for induced CRE and IFNB, they performed nearly as well or better than the more complex models we fit. Linear models trained on the multi-hit data are shown in Figs. 3.8B and 3.9B. Inspection revealed good qualitative correspondence with the sequence features described above. For example, the two CRE models show that an intact CREB site 1 is critical for maximizing the induced activity, while site 4 has the largest influence on the basal activity.

To quantify how well the linear models describe our experimental data, we com-

puted the correlation between their predictions and the observed activities for both the 261 single substitution variants in the independent single-hit data and the 27,000 variants in the multi-hit training sets. For the CRE, we found that the linear model for the induced state generates predictions that are highly correlated with the observed activities of both single- and multi-hit variants (r2 = 0.79, p ¡ 10-89 and r2 = 0.63, p ¡ 10-100, respectively). Remarkably, this model explains 90% of the biological variance in both data sets (compare to r2 = 0.89 and 0.67 between replicates, see above). The large number of multi-hit measurements ensures that this is not the result of over-fitting (r2 0.62 on five-fold cross-validation). In contrast, the induced IFNB model has a substantially poorer fit to the data (r2 = 0.61, p ¡ 10-54 and r2 = 0.071, p ¡ 10- 100, respectively). The difference in the fit of linear models appears to reflect the different architectures of the two enhancers. Most CRE multi-hit variants disrupt one or more of the non-overlapping consensus CREB sites, which caused large (median = 4.7-fold) and roughly additive reductions in its induced activity, until an apparent minimum is reached. Multiple substitutions were generally less detrimental to the induced IFNB enhancer (median decrease = 1.8-fold), which may reflect its initially weaker non-consensus binding sites or more complex interactions between its bound transcription factors.

Interactions between different positions in the enhancers can be modeled by including additional terms in the linear models. In initial efforts to model interactions, we trained models with terms that represented simple pair-wise interactions within and between the known binding sites. These terms captured statistical interactions between most binding sites and led to incremental improvements in the fit of the models (7% and 18% increases in r^2 for the induced CRE and IFNB enhancer models, respectively). It is evident, however, that more complex models would be required to fully explain the observed activities.

Because both enhancers showed evidence of nonlinear responses, we next at-

tempted to refine our models by incorporating functional nonlinearities. We fit a variety of models to the data, including ones describing either dinucleotide interactions or biophysical interactions between DNA-bound proteins. The best performing model was a simple 'linear-nonlinear' model. In this model, a sigmoidal transformation specified by parameters B and C is applied to the prediction of a linear model having parameters M_{bi} as defined above:

$$\log\ A(\sigma) = \log\left(B + C\frac{1}{1+\exp(\sum_{b,i} M_{bi}\sigma_{bi})}\right) \tag{3.15}$$

This type of model is widely used to describe systems where multiple inputs are combined to generate a response that interpolates monotonically, but not linearly, between minimum and maximum values. For the induced CRE data, this twoparameter nonlinearity increased r^2 by 16% as compared to the linear model. Because monotonic transformations have no effect on mutual information, this quantity was not meaningfully affected. Nevertheless, this linearnonlinear model has the virtue of being able to predict an upper limit to the expression level that can be achieved by reengineering the enhancer sequence.

Model parameters were optimized using regression. The optimal parameters for the linear part of this model are virtually identical ($r^2 = 0.98$) to the strictly linear model, but the two additional parameters that describe the sigmoidal nonlinearity allow the model to describe both minimum and maximum activation levels. Notably, this nonlinearity appears to capture much of the remaining nontechnical variance in the induced CRE data ($r^2 = 0.72$, $P < 10^{100}$, compared to $r^2 = 0.67$ between the two replicates). For the IFNB enhancer, the best performing models were those that incorporated dinucleotide interactions, which is consistent with its more complex architecture, although no model provided more than a modest improvement over the linear model (up to $r^2 = 0.10$, $P < 10^{100}$). Thus, although linear models are imperfect

representations of the underlying biological systems, in these cases they appear to provide a reasonable trade-off between complexity and predictive power.

Single-hit data

The single-hit scanning mutagenesis probes the effects of individual positions on the enchancers. We estimated the relative activity of each variant by comparing the median of its 13 mRNA/plasmid tag ratios to the median ratio for tags linked to the corresponding wild-type enhancer. We first focused on the CRE, which contains two consensus CREB dimer binding sites (denoted as sites 1 and 4 in Fig. 3.8A) separated by two monomer sites (sites 2 and 3). We found that 154 of the 261 possible single substitutions significantly altered its activity (5% FDR), with the majority (79%) resulting in decreased activity (Fig. 3.8B). The substitutions that resulted in the largest decreases were in or immediately flanking the CREB sites. Substitutions in the promoter-proximal CREB site 4 had the largest effects, which is consistent with reports of the cAMP responsiveness of CREB sites being inversely correlated with their distance from a TATA-box14. Within the two dimer sites, substitutions in the central CGs were the most deleterious. This is consistent with biochemical data that show that this dinucleotide is critical for high-affinity CREB-DNA interactions [65].

Substitutions at 47 of 61 positions outside of the CREB sites also caused significant (5% FDR), although generally more subtle, changes in activity. This may reflect the effects of cryptic non-CREB binding sites. In particular, two substitutions upstream of CREB site 1, as well as almost every substitu- tion in a C-rich motif flanking CREB site 4, resulted in increased CRE activity. These substitutions may therefore cause either increased recruitment of activating factors or decreased recruitment of repressors.

We next focused on the IFNB enhancer, which is a 44-nt sequence containing overlapping, nonconsensus binding sites for an ATF-2/c-Jun heterodimer, two IRF-3

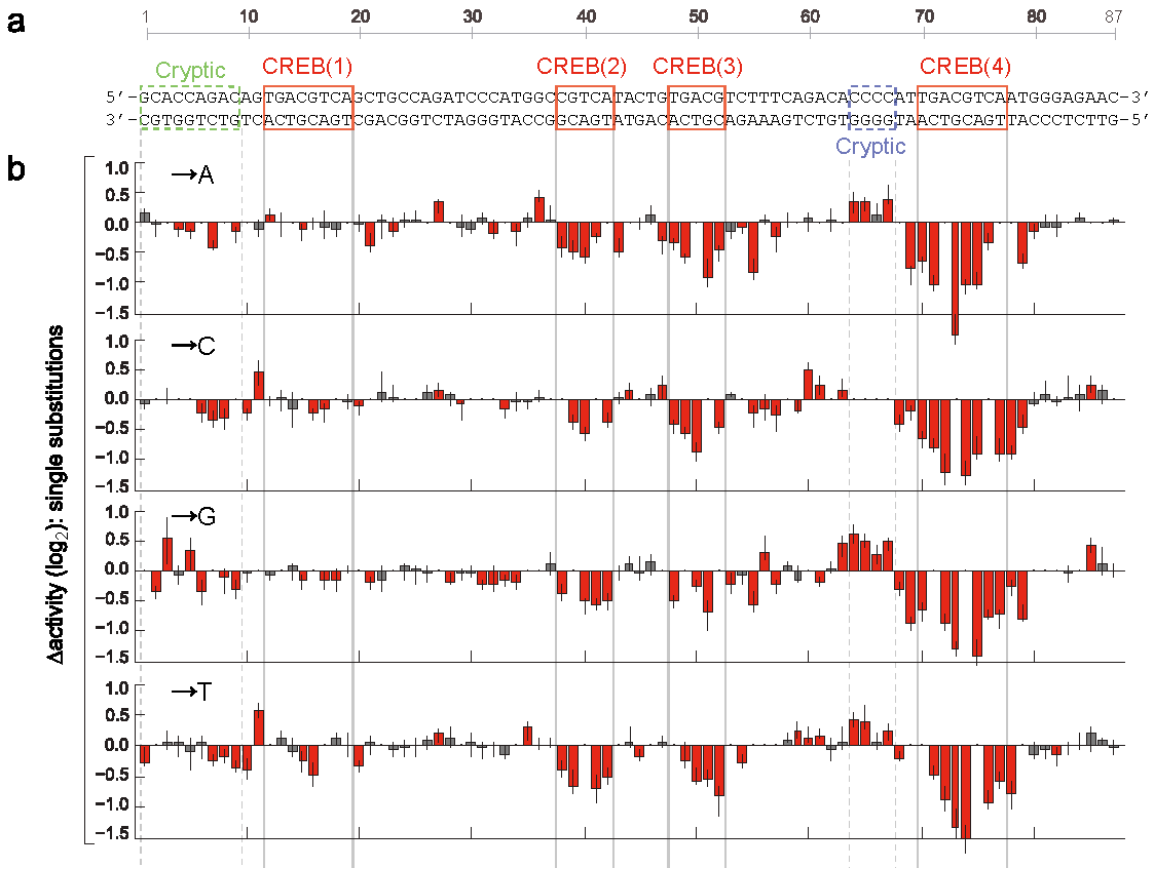

Figure 3.10: Single-hit scanning mutagenesis of the cAMP-responsive enhancer. (a) The CRE sequence with known and putative transcription factor binding sites indicated. (b) Changes in induced activity owing to single-nucleotide substitutions. Each bar shows the log-ratio of the median variant and wild-type activity estimates. Figure from [42].

and two IRF-7 proteins, and a p50/RELA (NF-B) heterodimer (Fig. 3.9A) [64]. We included a small amount of flanking genomic sequence, for a total length of 87 nt. We found that 83 of the 261 possible single substitutions altered the enhancers activity in virus-infected cells (5% FDR), and that almost all (92%) of these were within the 44-nt core (Fig. 3.9b).

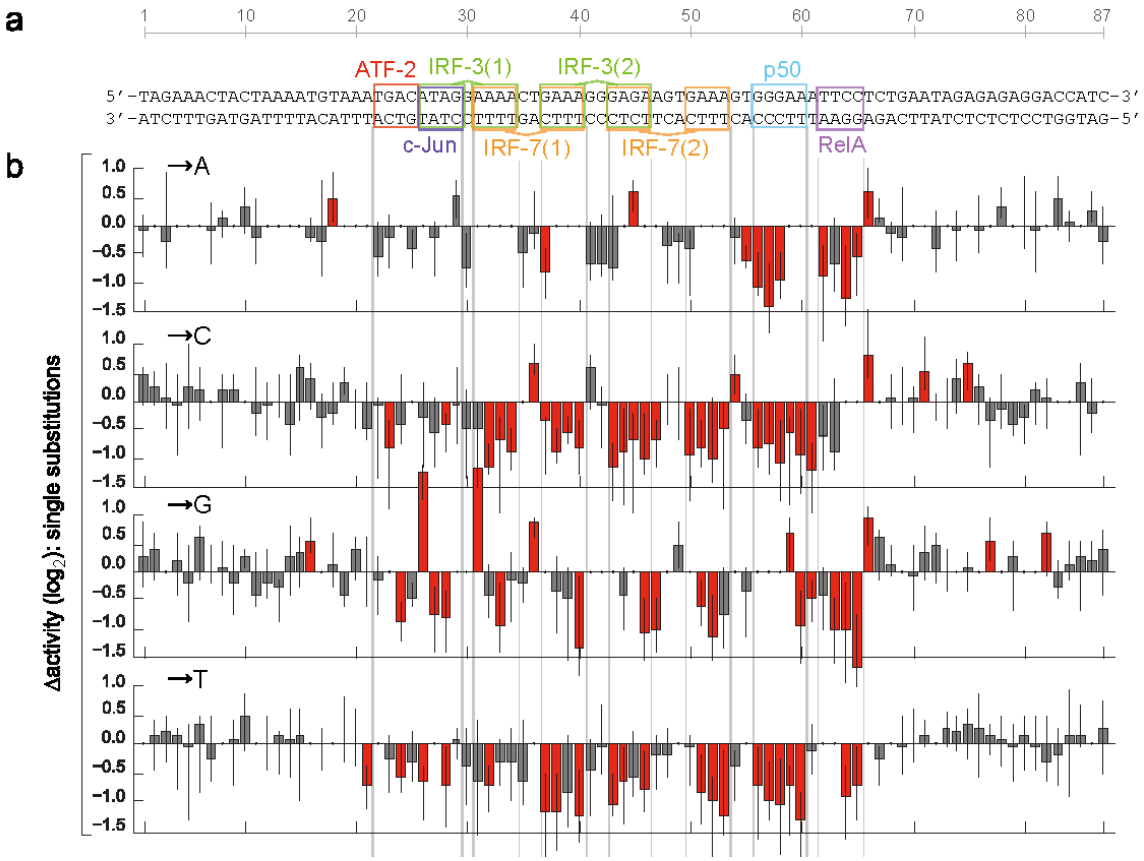

Figure 3.11: Single-hit scanning mutagenesis of the virus-inducible IFNB enhancer. (a) The IFNB enhancer with known transcription factor binding sites indicated. (b) Changes in induced activity owing to single-nucleotide substitutions. Each bar shows the log-ratio of the median variant and wild-type activity estimates. Figure from [42].

Optimization of enhancer sequences

To explore the potential for model-based optimization of synthetic regulatory elements in mammals, we next attempted to design enhancers with modified activities. We first attempted a greedy approach to maximize the induced enhancer activities. We selected, for each position, the nucleotide predicted to make the largest activity contribution according to the corresponding linear model. This resulted in changing the CRE at 36 of 87 positions (CRE-A1 in Fig. 3.12A). These changes left the consensus CREB sites intact, but introduced predicted activating mutations into the flanks of CREB sites 13 and into the two cryptic binding sites. For the IFNB enhancer, we limited modifications to the 44-nt core. This resulted in changes at 15 positions (IFNB- A1 in Fig. 3.13A), including conversion of every nonconsensus IRF half-site to the GAAA consensus and strengthening of the p50 half- site. We individually synthesized these two variants and then com- pared them to their wild types using a luciferase assay. We found that both new variants had significantly higher induced activities (2.1-fold for CRE-A1, P ¡ 0.0001, and 2.6-fold for IFNB-A1, P ¡ 0.0001; Figs. 3.12B,3.13B). Notably, the increase for CRE-A1 (2.1-fold) was substantially lower than predicted by the simple linear model (32-fold), but close to the value predicted by the linear-nonlinear model (1.7-fold). In contrast, the increase for IFNB-A1 (2.6-fold) was close to the value predicted by its linear model (2.1-fold). This difference likely reflects that the wild-type CRE is composed of consensus activator sites and therefore operates much closer to saturation than the IFNB enhancer. We also found, however, that both new variants had disproportionately higher uninduced activities (19-fold for CRE-A1 and 17-fold for IFNB-A1). This suggests that mutations that increase the induced activity of an enhancer may often decrease its inducibility, which would likely be detrimental in most biological and engineering contexts.

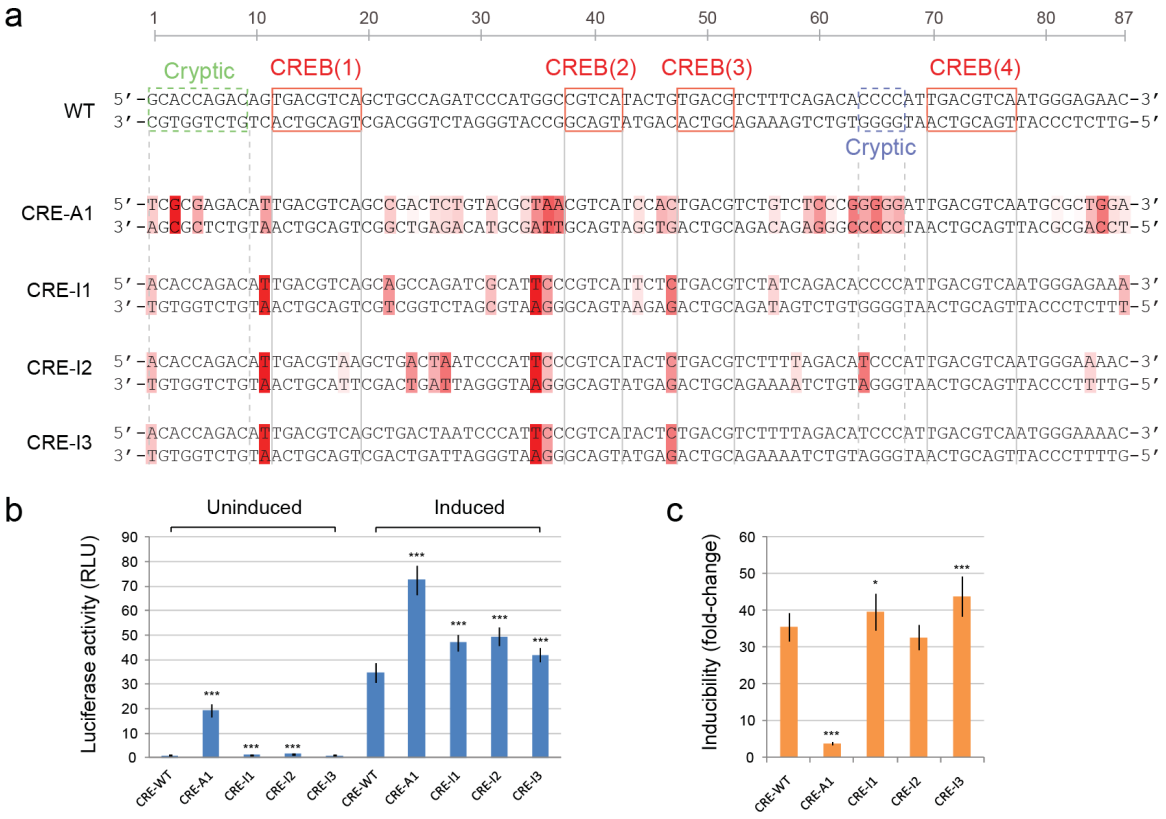

Figure 3.12: Model-based optimization of CRE. (a) CRE variants predicted to maximize induced activity (A1) or inducibility (I1-I3) based on linear models trained on multi-hit data. Differences from wild type are indicated by red shading. Darker shading indicates a higher predicted contribution to the change in activity. (b) Luciferase activity of the wild-type (WT) and optimized CRE variants in untreated and forskolin-treated cells. RLU, relative light unit. (c) Inducibility of the CRE variants in response to cAMP elevation caused by forskolin treatment. Blue bars show mean activity across 12 replicates in the induced or uninduced states. Error bars show s.e.m. (SE). All statistical comparisons are relative to WT in the same state; n.s., not significant; ***, $P \leq 0.0001$; two-tailed t-test. Orange bars show the ratio of the corresponding induced and uninduced mean activities. Error bars show the range from (induced mean induced SE)/(uninduced mean + uninduced SE) to (induced mean + induced SE)/(uninduced mean uninduced SE).Figure from [42].

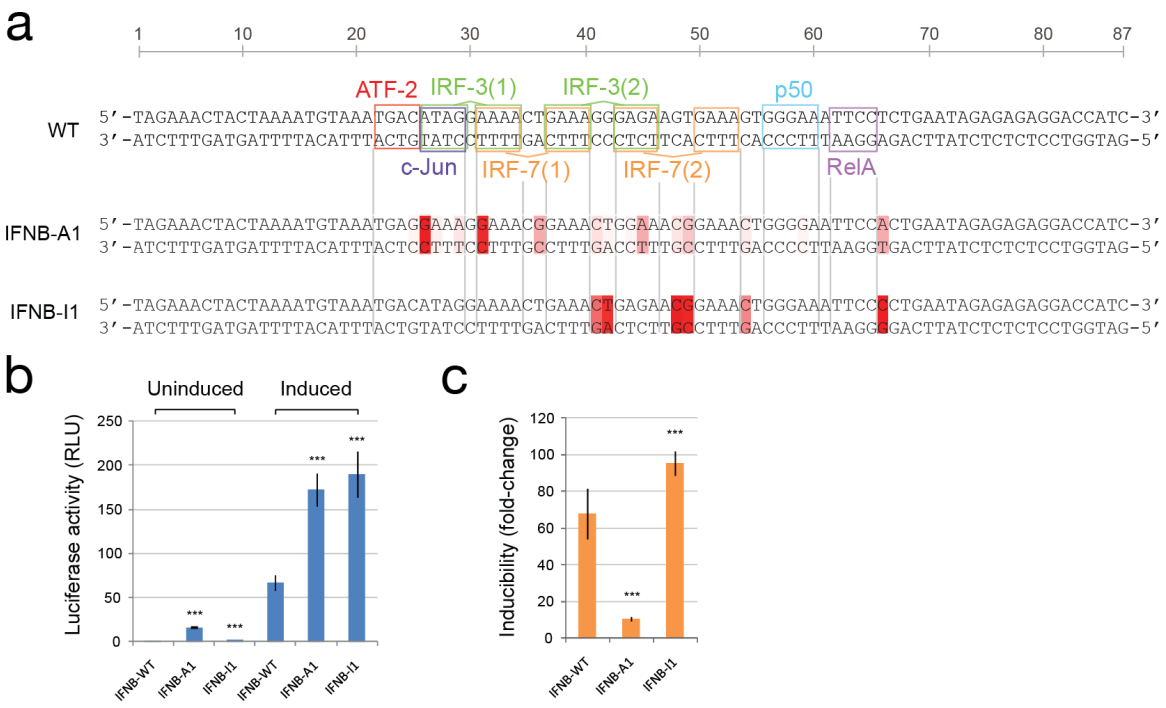

Figure 3.13: Model-based optimization of IFNB. (a) IFNB enhancer variants predicted to maximize induced activity (A1) or inducibility (I1) based on linear models trained on multi-hit data. (b) Luciferase activity of the WT and optimized IFNB enhancer variants in uninfected and virus-treated cells. (c) Inducibility of the IFNB enhancer variants in response to virus infection. Blue bars show mean activity across 12 replicates in the induced or uninduced states. Error bars show s.e.m. (SE). All statistical comparisons are relative to WT in the same state; n.s., not significant; ***, $P \leq 0.0001$; two-tailed t-test. Orange bars show the ratio of the corresponding induced and uninduced mean activities. Error bars show the range from (induced mean − induced SE)/(uninduced mean + uninduced SE) to (induced mean + induced SE)/(uninduced mean − uninduced SE). Figure from [42].

3.3.4 Summary

Massively parallel reporter assays of the kind we have used in our work enable functional analysis of transcriptional regulatory elements in cultured cells at significantly higher throughput than traditional bioluminescence- and fluorescence-based assays. The resulting data was used to map functional transcription factor binding sites at single-nucleotide resolution and to train quantitative sequence-activity models. This approach may help elucidate the biophysical basis of inducible and cell type-specific enhancer activity.

Beyond studying variants of naturally-occurring DNA sequences, the flexibility and decreasing cost of DNA synthesis is enabling construction and optimization of novel regulatory elements. Strong synthetic promoters have previously been selected from combinatorial libraries using FACS [66, 67]. It may be challenging, however, to design direct selection strategies for regulatory elements with more complex characteristics, such as optimal inducibility, dynamic range or cell type-specificity. Model-based optimization represents an alternative to direct selection. In this approach, all synthesized elements are first profiled in multiple cells states, with the resulting data being integrated to identify sequences that optimize complex objectives. This approach can be applied iteratively, which would be conceptually similar to genetic algorithm-based optimization [68]. With the development of more sophisticated mutagenesis and modeling strategies, we expect that this approach will provide a powerful tool for synthetic biology.

Chapter 4

Conclusions

In this thesis, we have described work on two different biological processes that rely directly on the information content of nucleic acid sequences. These functional sequences raise questions about both the mechanisms by which they exercise their influence, as well as the processes by which they are generated. In our work on regulatory sequences, we have investigated the first, while in our work on the immune system, we have investigated the second.

The adaptive immune system in vertebrates presents a remarkable internal implementation of the principles of natural selection. The dynamics of the immune cell repertoire in response to selective forces from pathogens might provide an ideal laboratory for understanding the working of evolution in general. Here, we have focused on inferring the 'neutral' model for this system by characterizing the diversity of immune cell receptors that can be generated by the molecular generation process. This is an essential first step for the analysis of the effects of selection on the repertoire.

With our 'neutral' model in hand, we now plan to analyze the functional receptor repertoires to look for deviations from the same. While it may be easy to identify statistical signatures of selection, it is much harder to interpret them in terms of function. We do not yet know how to parametrize receptor function and the selective

forces from pathogens to describe their dynamics. This is a challenging subject for future investigation.

The regulation of gene expression is implemented by a large diversity of complex mechanisms. These mechanisms need to be dissected for a full understanding of the relationship between genetic information and phenotypes. Here, we have investigated relatively simple instances of transcriptional regulation by modeling the sequence-function relationship for promoters and enhancers.

The general approach we have used, of perturbing the functional sequences and measuring their activities in large numbers, can be generalized to other systems where the precise sequence is critical for biological function. Future work could focus on building physically motivated models of the mechanisms of complex enhancers. Such models can also aid in the synthetic design of regulatory sequences, as we have demonstrated.

Besides the use of high-throughput sequencing to obtain large amounts of data, the two projects share the application of sophisticated statistical inference techniques to extract interpretable models from the data. The typical size of data sets from biological experiments is constantly increasing due to advances in multiple techologies. However, data by itself does not make biological research quantitative. It is essential to use model-based inference techniques to build quantitative understanding of biological systems.

References

[1] J. Shendure and H. Ji, "Next-generation DNA sequencing," *Nature Biotechnology* **26** (Oct., 2008) 1135–1145.

[2] E. Lander, "Initial impact of the sequencing of the human genome," *Nature* (Jan., 2011).

[3] G. J. Hannon, "RNA interference.," *Nature* **418** (July, 2002) 244–251.

[4] G. A. Maston, S. K. Evans, and M. R. Green, "Transcriptional regulatory elements in the human genome.," *Annual review of genomics and human genetics* **7** (2006) 29–59.

[5] C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J.-B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok, M. Kellis, P. A. Rolfe, K. T. Takusagawa, E. S. Lander, D. K. Gifford, E. Fraenkel, and R. A. Young, "Transcriptional regulatory code of a eukaryotic genome," *Nature* **431** (Sept., 2004) 99–104.

[6] G. A. T. McVean, "The Fine-Scale Structure of Recombination Rate Variation in the Human Genome," *Science* **304** (Apr., 2004) 581–584.

[7] E. R. Mardis, "The impact of next-generation sequencing technology on genetics," *Trends in Genetics* **24** (Mar., 2008) 133–141.

[8] C. Janeway, *Immunobiology.* the immune system in health and disease. Garland Pub, 2005.

[9] K. E. Lewis and D. O'Day, "Phagocytosis in Dictyostelium: Nibbling, Eating and Cannibalism," *Journal of Eukaryotic Microbiology* (1996).

[10] T. Farries, "Evolution of the complement system," *Immunology today* (1991).

[11] T. Boehm, "Design principles of adaptive immune systems," *Nature reviews. Immunology* **11** (Apr., 2011) 307–317.

[12] K. P. Murphy, P. Travers, M. Walport, and C. Janeway, *Janeway's immunobiology.* Garland Pub, 2008.

[13] M. P. Lefranc, V. Giudicelli, C. Ginestoux, J. Jabado-Michaloud, G. Folch, F. Bellahcene, Y. Wu, E. Gemrot, X. Brochet, J. Lane, L. Regnier, F. Ehrenmann, G. Lefranc, and P. Duroux, "IMGT(R), the international ImMunoGeneTics information system(R)," *Nucleic Acids Research* **37** (Jan., 2009) D1006–D1012.

[14] D. Jung and F. W. Alt, "Unraveling V(D)J Recombination: Insights into Gene Regulation," *Cell* **116** (Jan., 2004) 299–311.

[15] P. D. Hodgkin, W. R. Heath, and A. G. Baxter, "The clonal selection theory: 50 years since the revolution," *Nature immunology* **8** (Oct., 2007) 1019–1026.

[16] D. G. Schatz and P. C. Swanson, "V(D)J recombination: mechanisms of initiation.," *Annual review of genetics* **45** (2011) 167–202.

[17] N. Verkaik, R. Esveldt-van Lange, D. van Heemst, H. Bruggenwirth, J. Hoeijmakers, M. Zdzienicka, and D. C. van Gent, "Different types of V(D)J recombination and end-joining defects in DNA double-strand break repair

mutant mammalian cells," *European Journal of Immunology* **32** (2002), no. 3 701–709.

[18] M. R. Lieber, "The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End-Joining Pathway," *Annual Review of Biochemistry* **79** (June, 2010) 181–211.

[19] M. R. Lieber and T. E. Wilson, "SnapShot: Nonhomologous DNA end joining (NHEJ).," *Cell* **142** (Aug., 2010) 496–496.e1.

[20] J. J. Lafaille, A. DeCloux, M. Bonneville, Y. Takagaki, and S. Tonegawa, "Junctional sequences of T cell receptor gamma delta genes: implications for gamma delta T cell lineages and for a novel intermediate of V-(D)-J joining.," *Cell* **59** (Dec., 1989) 859–870.

[21] M. R. Lieber, H. Lu, J. Gu, and K. Schwarz, "Flexibility in the order of action and in the enzymology of the nuclease, polymerases, and ligase of vertebrate non-homologous DNA end joining: relevance to cancer, aging, and the immune system.," *Cell Research* **18** (Jan., 2008) 125–133.

[22] M. Y. Monod, V. Giudicelli, D. Chaume, and M. P. Lefranc, "IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs," *Bioinformatics* **20** (July, 2004) i379–i385.

[23] A. M. Sherwood, C. Desmarais, R. J. Livingston, J. Andriesen, M. Haussler, C. S. Carlson, and H. Robins, "Deep Sequencing of the Human TCR and TCR Repertoires Suggests that TCR Rearranges After and T Cell Commitment," *Science translational medicine* **3** (July, 2011) 90ra61–90ra61.

[24] J. D. Freeman, R. L. Warren, J. R. Webb, B. H. Nelson, and R. A. Holt, "Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing.," *Genome Research* **19** (Oct., 2009) 1817–1824.

[25] N. Jiang, J. A. Weinstein, L. Penland, R. A. White, D. S. Fisher, and S. R. Quake, "Determinism and stochasticity during maturation of the zebrafish antibody repertoire.," *Proceedings of the National Academy of Sciences* **108** (2011), no. 13 5348–5353.

[26] H. Robins, P. V. Campregher, S. K. Srivastava, A. Wacher, C. J. Turtle, O. Kahsai, S. R. Riddell, E. H. Warren, and C. S. Carlson, "Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells.," *Blood* **114** (Nov., 2009) 4099–4107.

[27] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions (Wiley Series in Probability and Statistics)*. Wiley-Interscience, 2 ed., 2008.

[28] Y. Wang, K. J. Jackson, B. Gäeta, W. Pomat, P. Siba, W. A. Sewell, and A. M. Collins, "Genomic screening by 454 pyrosequencing identifies a new human IGHV gene and sixteen other new IGHV allelic variants.," *Immunogenetics* **63** (2011), no. 5 259–265.

[29] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)* **39** (1977), no. 1 1–38.

[30] M. E. Wallace, M. Bryden, S. C. Cose, R. M. Coles, T. N. Schumacher, A. Brooks, and F. R. Carbone, "Junctional biases in the naive TCR repertoire control the CTL response to an immunodominant determinant of HSV-1.," *Immunity* **12** (May, 2000) 547–556.

[31] H. Robins, S. K. Srivastava, P. V. Campregher, C. J. Turtle, J. Andriesen, S. R. Riddell, C. S. Carlson, and E. H. Warren, "Overlap and effective size of the human CD8+ T cell receptor repertoire.," *Science translational medicine* **2** (Sept., 2010) 47ra64.

[32] G. H. Gauss and M. R. Lieber, "Mechanistic constraints on diversity in human V(D)J recombination.," *Molecular and Cellular Biology* **16** (Jan., 1996) 258–269.

[33] K. S. Blum and R. Pabst, "Lymphocyte numbers and subsets in the human blood - Do they mirror the situation in all organs?," *Immunology Letters* **108** (2007), no. 1 45–51.

[34] D. Mason, "A very high level of crossreactivity is an essential feature of the T-cell receptor.," *Immunology today* **19** (Sept., 1998) 395–404.

[35] J. P. Cabaniols, N. Fazilleau, A. Casrouge, P. Kourilsky, and J. M. Kanellopoulos, "Most alpha/beta T cell receptor diversity is due to terminal deoxynucleotidyl transferase.," *The Journal of experimental medicine* **194** (2001), no. 9 1385–1390.

[36] M. F. Quigley, H. Y. Greenaway, V. Venturi, R. Lindsay, K. M. Quinn, R. A. Seder, D. C. Douek, M. P. Davenport, and D. A. Price, "Convergent recombination shapes the clonotypic landscape of the naive T-cell repertoire.," *Proceedings of the National Academy of Sciences of the United States of America* **107** (Nov., 2010) 19414–19419.

[37] V. Venturi, M. F. Quigley, H. Y. Greenaway, P. C. Ng, Z. S. Ende, T. McIntosh, T. E. Asher, J. R. Almeida, S. Levy, D. A. Price, M. P. Davenport, and D. C. Douek, "A mechanism for TCR sharing between T cell

subsets and individuals revealed by pyrosequencing.," *Journal of immunology (Baltimore, Md. : 1950)* **186** (Apr., 2011) 4285–4294.

[38] V. Venturi, D. A. Price, D. C. Douek, and M. P. Davenport, "The molecular basis for public T-cell responses?," *Nature reviews. Immunology* **8** (Mar., 2008) 231–238.

[39] T. Mora, A. M. Walczak, W. Bialek, and C. G. Callan, "Maximum entropy models for antibody diversity.," *Proceedings of the National Academy of Sciences of the United States of America* **107** (Mar., 2010) 5405–5410.

[40] S. B. Carroll, "Endless Forms," *Cell* **101** (June, 2000) 577–580.

[41] J. B. Kinney, A. Murugan, C. G. Callan, and E. C. Cox, "Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence.," *Proceedings of the National Academy of Sciences of the United States of America* **107** (May, 2010) 9158–9163.

[42] A. Melnikov, A. Murugan, X. Zhang, T. Tesileanu, L. Wang, P. Rogov, S. Feizi, A. Gnirke, C. G. Callan, J. B. Kinney, M. Kellis, E. S. Lander, and T. S. Mikkelsen, "Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay," *Nature Biotechnology* **30** (Feb., 2012) 271–277.

[43] U. Gerland, J. D. Moroz, and T. Hwa, "Physical constraints and functional characteristics of transcription factor-DNA interaction," *Proceedings of the National Academy of Sciences of the United States of America* **99** (Sept., 2002) 12015–12020.

[44] B. Müller-Hill, *The Lac Operon.* A Short History of a Genetic Paradigm. De Gruyter, 1996.

[45] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips, "Transcriptional regulation by the numbers: models.," *Current opinion in genetics & development* **15** (Apr., 2005) 116–124.

[46] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, T. Kuhlman, and R. Phillips, "Transcriptional regulation by the numbers: applications.," *Current opinion in genetics & development* **15** (Apr., 2005) 125–135.

[47] T. Kaplan, X.-Y. Li, P. J. Sabo, S. Thomas, J. A. Stamatoyannopoulos, M. D. Biggin, and M. B. Eisen, "Quantitative Models of the Mechanisms That Control Genome-Wide Patterns of Transcription Factor Binding during Early Drosophila Development," *PLoS Genetics* **7** (2011), no. 2 –.

[48] U. Alon, "An introduction to systems biology: design principles of biological circuits,".

[49] T. Kuhlman, Z. Zhang, M. Saier, and T. Hwa, "Combinatorial transcriptional control of the lactose operon of Escherichia coli," *Proceedings of the National Academy of Sciences* **104** (2007), no. 14 6043.

[50] C. E. Shannon and W. Weaver, "The mathematical theory of communication," 1964.

[51] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, July, 2006.

[52] J. B. Kinney, G. Tkačik, and C. G. Callan, "Precise physical models of protein–DNA interaction from high-throughput data," *Proceedings of the National Academy of Sciences* **104** (2007), no. 2 501.

[53] D. Earl and M. Deem, "Parallel tempering: Theory, applications, and new perspectives," *Physical Chemistry Chemical Physics* **7** (2005), no. 23 3910–3916.

[54] O. G. Berg and P. H. von Hippel, "Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters," *Journal of Molecular Biology* **193** (Feb., 1987) 723–750.

[55] Y. Takeda, A. Sarai, and V. M. Rivera, "Analysis of the sequence-specific interactions between Cro repressor and operator DNA by systematic base substitution experiments," *Proceedings of the National Academy of Sciences of the United States of America* **86** (Jan., 1989) 439–443.

[56] W. S. Reznikoff, "The lactose operon-controlling elements: a complex paradigm," *Molecular Microbiology* **6** (Sept., 1992) 2419–2422.

[57] W. Ross, K. K. Gosink, J. Salomon, K. Igarashi, C. Zou, A. Ishihama, K. Severinov, and R. L. Gourse, "A third recognition element in bacterial promoters: DNA binding by the alpha subunit of RNA polymerase.," *Science* **262** (Nov., 1993) 1407–1413.

[58] Y. Zhao, D. Granas, and G. D. Stormo, "Inferring binding energies from selected binding sites," *PLoS Comput Biol* **5** (Dec., 2009) e1000590.

[59] R. P. Patwardhan, C. Lee, O. Litvin, D. L. Young, D. Pe'er, and J. Shendure, "High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis," *Nature Biotechnology* **27** (Jan., 2009) 1173–1175.

[60] C. Beisel and R. Paro, "Silencing chromatin: comparing modes and mechanisms," *Nature Reviews Genetics* **12** (Feb., 2011) 123–135.

[61] K. A. Zawadzki, A. V. Morozov, and J. R. Broach, "Chromatin-dependent transcription factor accessibility rather than nucleosome remodeling

predominates during global transcriptional restructuring in Saccharomyces cerevisiae," *Molecular biology of the cell* **20** (Aug., 2009) 3503–3513.

[62] L. A. Mirny, "Nucleosome-mediated cooperativity between transcription factors," *Proceedings of the National Academy of Sciences of the United States of America* (Dec., 2010).

[63] F. Fan and K. V. Wood, "Bioluminescent Assays for High-Throughput Screening," *ASSAY and Drug Development Technologies* **5** (Feb., 2007) 127–136.

[64] D. Panne, T. Maniatis, and S. C. Harrison, "An Atomic Model of the Interferon- Enhanceosome," *Cell* **129** (June, 2007) 1111–1123.

[65] D. Benbrook and N. Jones, "Different Binding Specificities and Transactivation of Variant Cres by Creb Complexes," *Nucleic Acids Research* **22** (1994), no. 8 1463–1469.

[66] B. E. Bernstein, J. A. Stamatoyannopoulos, J. F. Costello, B. Ren, A. Milosavljevic, A. Meissner, M. Kellis, M. A. Marra, A. L. Beaudet, J. R. Ecker, P. J. Farnham, M. Hirst, E. S. Lander, T. S. Mikkelsen, and J. A. Thomson, "The NIH Roadmap Epigenomics Mapping Consortium," *Nature Biotechnology* **28** (Oct., 2010) 1045–1048.

[67] G. M. Edelman, R. Meech, G. C. Owens, and F. S. Jones, "Synthetic promoter elements obtained by nucleotide sequence variation and selection for activity,".

[68] M. R. Schlabach, J. K. Hu, M. Li, and S. J. Elledge, "Synthetic design of strong promoters," *Proceedings of the National Academy of Sciences of the United States of America* **107** (Feb., 2010) 2538–2543.