

Annotation and exploration of the metazoan regulatory genome



"We think it has something to do with your genome."

Marc S. Halfon, Ph.D.

Departments of Biochemistry, Biological Sciences, and Biomedical Informatics, SUNY at Buffalo NY State Center of Excellence in Bioinformatics & Life Sciences Department of Molecular and Cellular Biology, Roswell Park Cancer Institute

http://halfonlab.ccr.buffalo.edu mshalfon@buffalo.edu

Genomes are easy...

Let's learn the alphabet!



http://www.newlybaby.com

...or are they?
From human chr. 16
(<6.4 kb of 3×10^6 kb, 0.0002%)

We know the sequence—but can we understand it?
From human chr. 16
(<6.4 kb of 3×10^6 kb, 0.0002%)

Understanding the genome

We don't know (all) the language:

Гостиня Анна Павловна начала понемногу наполняться. Приехала высшая знать Петербурга, люди самые разнородные по возрасту и характеру, но одинаковые по обществу, в каком все жили; приехала дочь князя Василия, красавица Элен, захватившая за отцом, чтобы с ним вместе ехать на праздник сепьянина. Она была в шифре и бальном платье. Приехала и известная, как la femme la plus séduisante de Pétersbourg 1, молодая, маленькая княгиня Болконская, прощлую зиму вышедшая замуж и теперь не выходящая в большой свет по причине своей беременности, но едущая еще на небольшие вечера. Приехал князь Николит, сын князя Василия, с Мортемаром, которого он предвещал; князь и аббат Морю и многие другие.

— Вы не видели еще, — или: — вы не знакомы с ma tante? — говорила Анна Павловна приезжавшим гостям и весьма серьезно подолвила их к маленькой старушке в высоких бантах, выплывшей из другой комнаты, как скоро стали приехать гости, называла их по имени, медленно переводя глаза с гостя на ma tante, и потом отходила.

Все гости совершали обряд приветствования никому не известной, никому не интересной и не нужной тетюшке. Анна Павловна в грустном, торжественном участии следила за их приветствиями, молчаливо одобряя их. Ma tante каждому говорила в одних и тех же выражениях о его здоровье, о своем здоровье и о здоровье ее величества, которое нынче было, слава Богу, лучше. Все подходившие, из принципа не выказывая поспешности, с чувством облегчения исполненной тяжелой обязанности отходили от старушки, чтоб уж весь вечер ни

Understanding the genome

Even if we did, we don't know the grammar or punctuation:

annapavlovnasdrawingroomwasgraduallyfillingthehighestpetersburgsocietywasassembled herepeopleidifferingwidelyinageandcharacterbutallinthesocialcircletowhichtheybelonged princevasiliididgatherthebestofthelencametoakcherfathertheambassadorcountermenin theworealldressandherbadgesmaidofhonortheyouthfullittleprincessbolkonskayaknowna slafemmetaplusseduisantedepetersbourgwasalsothereshehadbeenmarriedduringtheprevious winterandbeingpregnantdidnotgotanylargegatheringsbutonlytoasmallreceptionsprincevasil ssonhipolytheadcomewithmortemarwhomehintroductedtheabbeurionandmanyothershadla socmetoeachnewarrivannapavlovnaidyaouhaveenjoyetmesiamourionduonodnotknowmya untandverygravelyconductedhimorthertoitleoldladywearinglargebowsofribboninhercapw hadohomesailinginfromanotherroomassonastheguestsbeganoutingandslowlyturningher yesfromthevisitorsheantannapavlovnamentionedeachonesnameandthenlfttheachvisit orperformedtheceremonyofgreetingthistoldantwhomnotionoftheknewntionofthemyant edoknonandnotoneofthecareaboutannapavlovnaobservedithegreetingwithmournfulm doleminterestandsilentapprovaltheauntspokeofeachofthimthemasameswordsbouttheirhealt handherowndandthehealthofhermajestywhohankogdowasbettertodayandeachvisitorwhopol itenesspreventedhishowingimpatienceleftthelddwomanwithenseseofleicfatthavingperform edavexatiousdutyanddidnotreturntohertheholeeveningtheyoungprincessbolkonskayahadr oughtsomeworkingold.

1428 "nucleotides"

Understanding the genome

And even then, do we understand what it means?

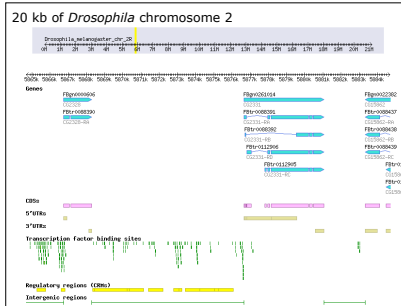
Anna Pavlovna's drawing room was gradually filling. The highest Petersburg society was assembled there: people differing widely in age and character but alike in the social circle to which they belonged. Prince Vasili's daughter, the beautiful Helene, came to take her father to the ambassador's entertainment; she wore a ball dress and her badge as maid of honor. The youthful little Princess Bolkonskaya, known as la femme la plus seduisante de Petersbourg, was also there. She had been married during the previous winter, and being pregnant did not go to any large gatherings, but only to small receptions. Prince Vasili's son, Hippolyte, had come with Mortemart, whom he introduced. The Abbe Morio and many others had also come.

To each new arrival Anna Pavlovna said, "You have not yet seen my aunt," or "You do not know my aunt?" and very gravely conducted him or her to a little old lady, wearing large bows of ribbon in her cap, who had come sailing in from another room as soon as the guests began to arrive, and slowly turning her eyes from the visitor to her aunt, Anna Pavlovna mentioned each one's name and then left them.

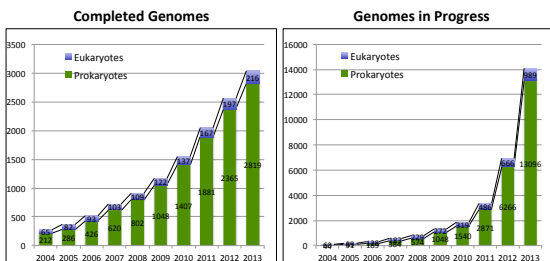
—Tolstoy, War and Peace

Genome Annotation

Having the raw genome sequence is therefore just a beginning. To make use of the sequence, we must **annotate** it—describe the function of each basepair of sequence.



A lot of genomes have been sequenced



<i>Buchnera aphidicola</i> Bc (bacterium)	422 kb	<i>Oryza sativa</i> (rice)	420,000 kb
<i>Bacillus anthracis</i> (anthrax)	5228 kb	<i>Mus musculus</i> (mouse)	2,493,000 kb
<i>S. cerevisiae</i> (yeast)	12,069 kb	<i>Homo sapiens</i> (human)	2,900,000 kb
<i>Drosophila melanogaster</i> (fruit fly)	137,000 kb	<i>Amoeba dubia</i> (amoeba)	670,000,000 kb

Source: ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/

so what?

Genome sequencing helps in:

- identifying new genes (“gene discovery”)
- identifying mutations and variations
- looking at chromosome organization and structure
- finding gene regulatory sequences
- comparative genomics

These in turn lead to advances in:

- medicine
- agriculture
- biotechnology
- understanding evolution and other basic science questions

What’s in a genome?

Genes (protein coding)

But... less than 2% of the human genome encodes proteins

Genes (non-protein coding: rRNA, tRNA, miRNAs, etc.)

Other than genes, what is there?

- structural sequences (scaffold attachment regions)
- regulatory sequences
- other (including transposons, retroviral insertions, etc.)

Protein-coding genes, Non-protein-coding genes

- Genes are easier to find than other functional elements
- Why?
 - Genes are transcribed—which means that we can identify them by looking at RNA
 - traditionally this has been done by cDNA or EST sequencing, more recently by microarray, SAGE, next-gen sequencing, etc.

Gene prediction

- We can also find (predict) genes using computational methods
- For example protein-coding genes have recognizable features
 - open reading frames (ORFs)
 - codon bias
 - known transcription and translational start and stop motifs (promoters, 3' poly-A sites)
 - splice consensus sequences at intron-exon boundaries
- We can design software to scan the genome and identify these features
- Some of these programs work quite well, especially in bacteria and simpler eukaryotes with smaller and more compact genomes

Validating predictions and refining gene models

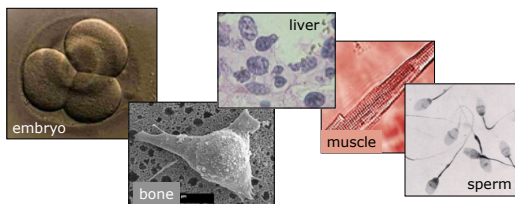
- Standard types of evidence for validation of predictions include:
- confidence** ↓
- match to previously annotated cDNA
 - match to EST from same organism
 - similarity of nucleotide or conceptually translated protein sequence to sequences in GenBank
 - protein structure prediction match to a PFAM domain
 - associated with recognized promoter sequences, i.e. TATA box, CpG island
 - known phenotype from mutation of the locus

What's in a genome?

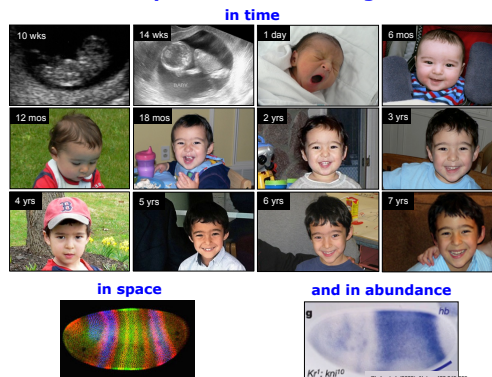
- Genes** (protein coding and non-coding)
- But . . . only <2% of the human genome encodes proteins
- Other than genes, what is there?**
- structural sequences (scaffold attachment regions)
 - **regulatory sequences**
 - other (including transposons, retroviral insertions, etc.)

Every cell has the same DNA and therefore the same genes.

But different genes need to be "on" and "off" in different types of cells. Therefore, gene expression must be *regulated*.



Gene expression must be regulated



Importance of gene regulation

evolution

common variation

pattern

behavior

chromosome inactivation

metabolism

pathology (mutation)

We commonly think of mutation as a change in a protein sequence, e.g., sickle cell anemia:

```

[wild type allele] >gi|28302128|ref|NM_000518.4| Homo sapiens hemoglobin, beta (HBB), mRNA
ATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTACTGCCCTGTGGGGCAAGGTGAACGTGATGAAGTTGGTGAAGC
CTGGGGCAGGCTGCTG...

>gi|4504349|ref|NP_000509.1| beta globin [Homo sapiens]
MVHLTPEVKSAVTALMGKVNVDVEVGGELGRLL...

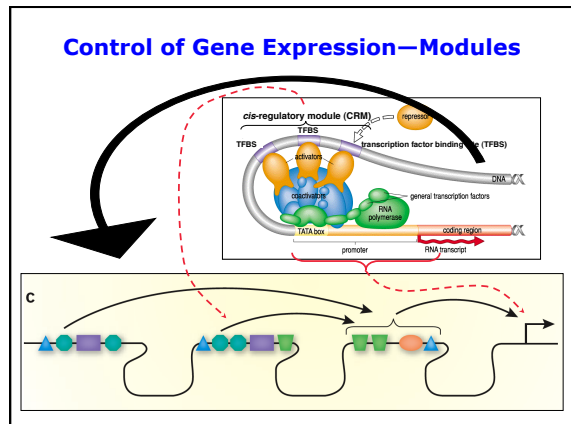
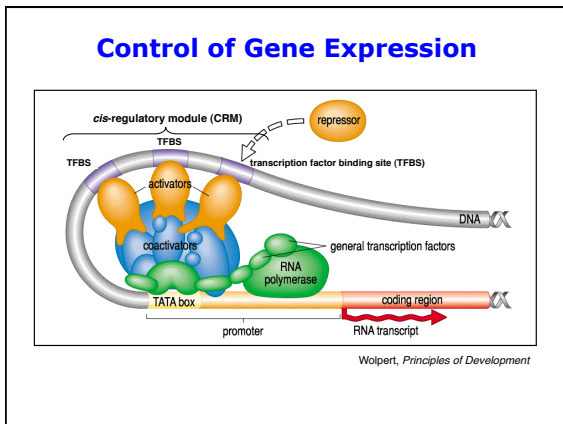
[sickle cell allele]
ATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTACTGCCCTGTGGGGCAAGGTGAACGTGATGAAGTTGGTGAAGC
CTGGGGCAGGCTGCTG...

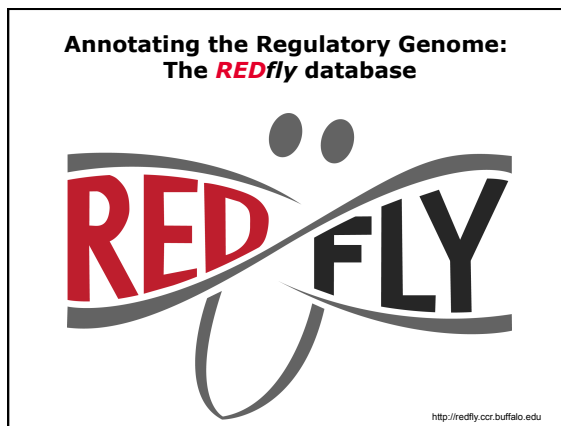
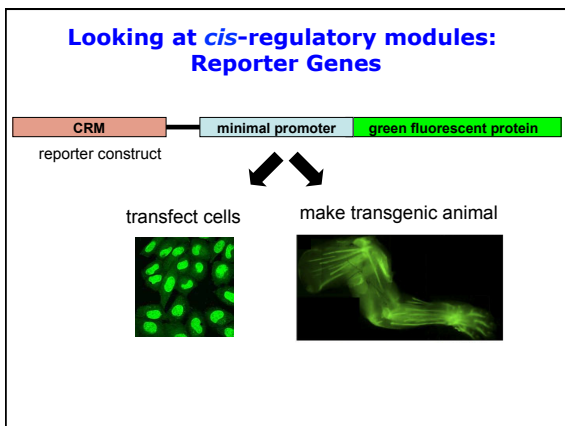
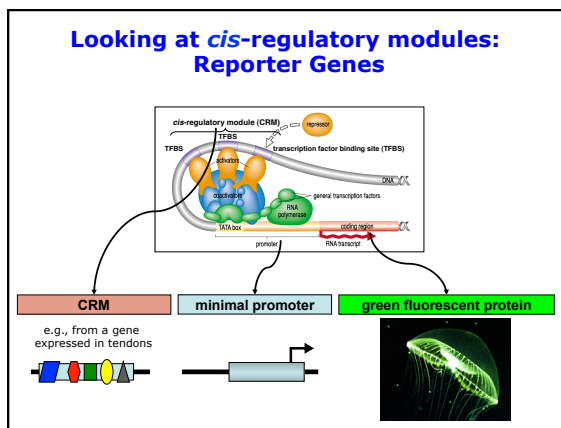
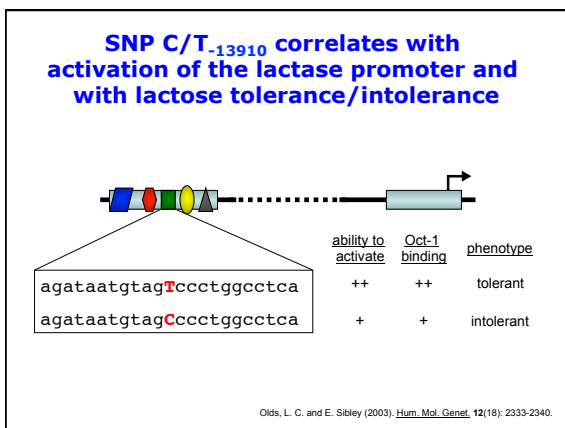
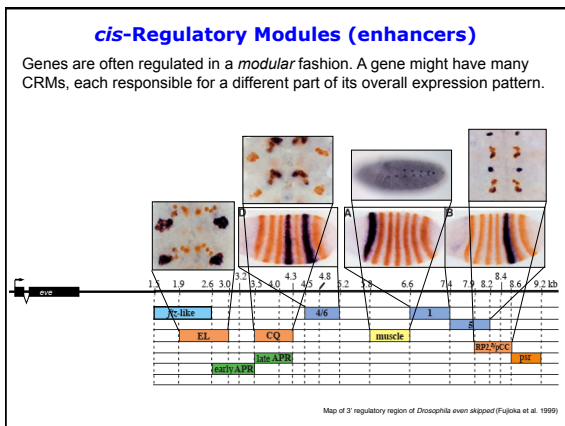
MVHLTVEKSAVTALMGKVNVDVEVGGELGRLL...
    
```

But mutations can also affect non-protein coding genes and gene regulatory regions.

However, these are much harder to detect and there are not that many known examples in humans (we'll see one in a little bit).

When gene regulation goes awry...





RED FLY Regulatory Element Database for Drosophila
 Database of known *Drosophila cis*-regulatory elements

- v4.0 just released, all sequences now R6 coordinates
- over 5500 CRMs so far from >500 genes
- based on >11,600 reporter gene assays (>11,000 transgenic in vivo)
- over 2000 TF binding sites (footprint & EMSA)
- includes sequence and expression pattern**
- includes graphical views and links to other relevant databases
- average ~400 hits/month, >100 citations (ISI)

<http://redfly.ccr.buffalo.edu> @REDfly_database



We can test assumptions using the REDfly collection

280 non-overlapping CRMs:

Li, Zhu, He, Sinha, and Halfon. (2007). *Genome Biol.*, 8:R101

SCRMshaw: Supervised CRM discovery

Kantorovitz et al. (2009). *Dev. Cell.*, 17:568-579
 Kazemian et al. (2011). *Nuc. Acids Res.* 39:9463-9472
 Kazemian et al. (2014). *Genome Biol. Evol.* 6:2301-2320

Halfon lab (SUNY Buffalo), Saurabh Sinha lab (UI Urbana-Champaign)

Supervised motif-blind approaches

We can search for genomic windows with word profiles similar to those in a *training set* of known CRMs. We do not concern ourselves with trying to discover the identities of the most relevant words at this stage.

Get set of similar CRMs from REDfly → Get *k*-mer profiles and compute score → Search genome for high-scoring windows

$$S = \sum_{w \in W} z(w) n(w)$$

with Saurabh Sinha, UI Urbana-Champaign

SCRMshaw has a high success rate

In vivo validation:
~80% true-positive rate

In silico validation:
83% recovery rate ($P \approx 0$)

Kantorovitz et al. (2009). *Dev. Cell.*, 17:568-579; Kazemian et al. (2011). *Nuc. Acids Res.* 39:9463-9472; unpublished data.

Discovering new regulatory regions

In mouse

LYL1 Promoter C1ORF164 EBF3

2/2 successful predictions (100%)

use of additional post-search filtering enhances tissue-specificity of results

Kantorovitz, M.R., Kazemian, ..., Halfon, M.S. and Sinha, S. (2009). *Dev. Cell.* 17:568-579.

Cross-species SCRMshaw: Can *Drosophila* help us discover CRMs in other insects?

1. Train on *Drosophila* CRMs as previously
2. Make predictions in other genomes: *Anopheles gambiae*, *Nasonia vitripennis*, *Apis mellifera*, *Tribolium castaneum*
3. Test high-scoring sequences in transgenic *Drosophila*

Kazemian et al. (2014) *Genome Biology and Evolution*

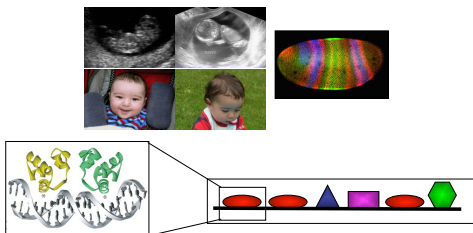
This information would have been impossible to obtain without the kind of large-scale study enabled by the *REDfly* database.

These data will help us to understand:

- how regulatory modules are organized
- how to identify regulatory modules by examining the genome sequence (and therefore also *mutations* in regulatory sequences)
- how genes are regulated at different times and places
- how gene regulation has evolved

Why bother?

Ultimately, we'd like to be able to describe all of development in terms of gene expression and regulation. That is, in every cell, at every time, which genes are on or off, and why.



And help to understand how we go from



here . . .

. . . to here . . .

. . . to here!