Bioinformatic approaches to discover post-transcriptional regulatory elements in human mRNAs

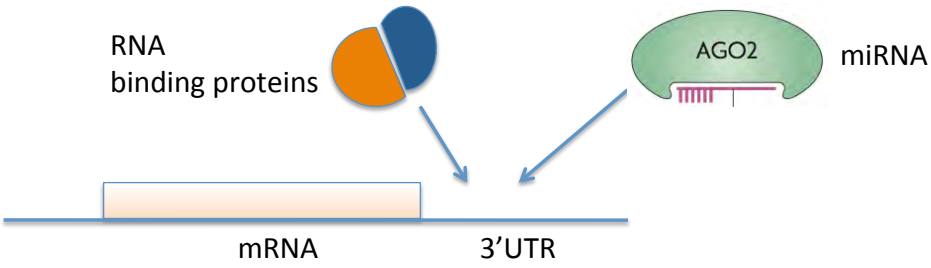
Chris Brown

Bioinformatics mainly: Ambarish Biswas Brad Croft, Gareth Gillard.

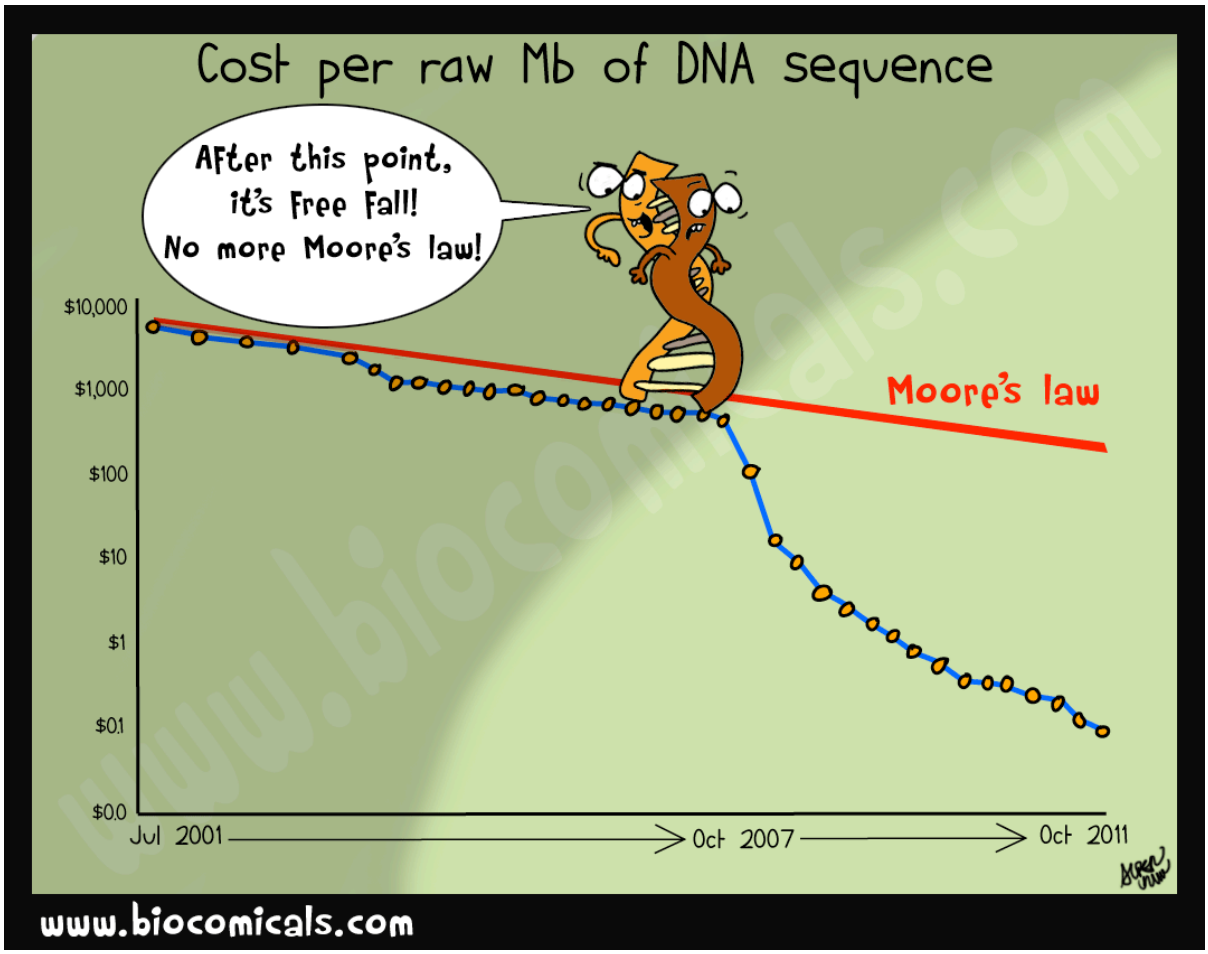
Cell Biology mainly: Sam Tayler-Wardell, Lim Chen

Biochemistry and Genetics Otago

University of Otago



Post-transcriptional control through the 3'UTR



GIZMODO UK

Join Login

UK NEWS GADGETS DESIGN WATCH THIS WTF SCIENCE APPLE ANDROID

SCIENCE October 3, 2013 - 5:00pm

"Build-a-Baby" DNA Database Patent Sounds Worryingly Like *Gattaca*

Gerald Lynch - Do you remember Gattaca, the 1997 dystopian sci-fi film in which Ethan Hawke endures a lifetime of prejudice because his parents chose not to supercharge his genes at birth? It's an unsettling vision of an intolerant future, and one that could become a reality if 23andMe's "Build-a-Baby" DNA database patent ever becomes an actual thing.

The "Family Traits Inheritor Calculator" could be used to predict the chance of a child being born with an inherited disease, as well as picking up details of a person's future appearance, such height and weight, and maybe even personality.

And though 23andMe has since distanced itself from the potential use of the tool as a designer baby catalogue for fertility clinics, experts fear that this could be the technology's ultimate application.

Gattaca

Andrew Niccol

Screenwriter

Andrew M. Niccol is a New Zealand screenwriter, producer, and director. He wrote and directed Gattaca, S1m0ne, In Time, and Lord of War. Wikipedia

Born: January 1, 1960, Paraparaumu

Spouse: Rachel Roberts (m. 2002)

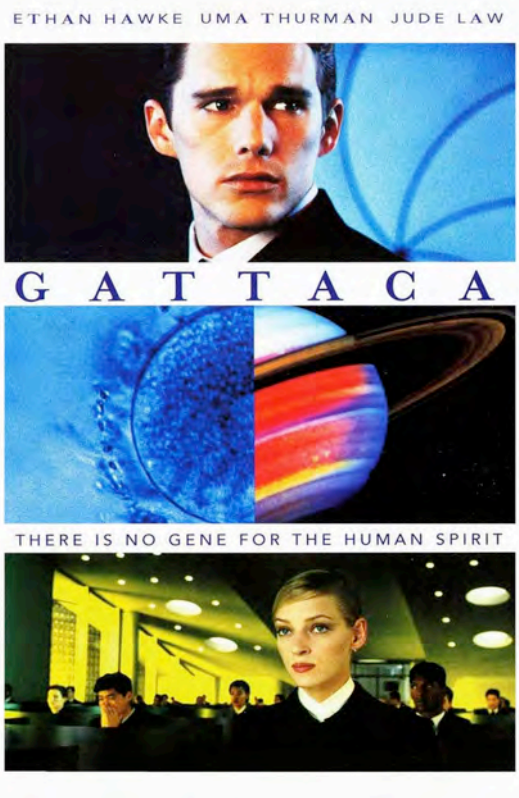
Awards: Saturn Award, Saturn Award for Best Writing, More

Children: Jack Niccol, Ava Lila Rae Niccol, Mia Niccol

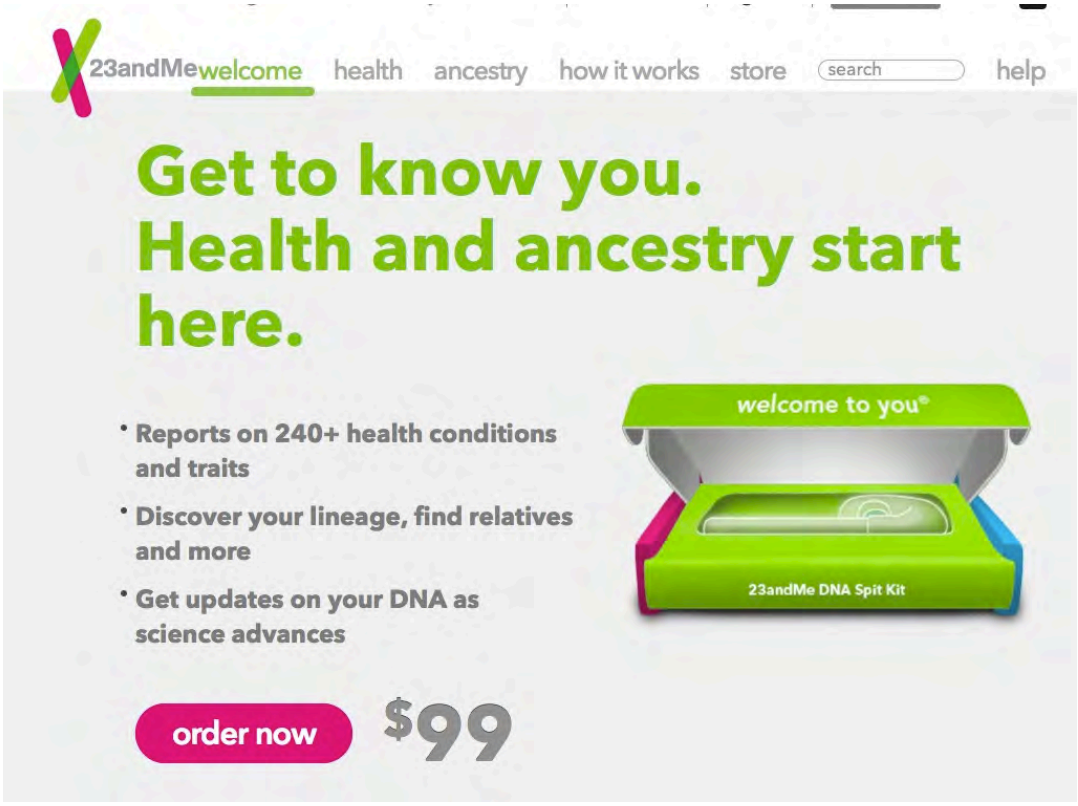
Siblings: Linda Niccol

Movies

The Host 2013 | In Time 2011 | Gattaca 1997 | Lord of War 2005 | The Truman S... 1998



https://www.youtube.com/watch?v=jpm4T2-v_bw



Forbes · New Posts · Most Popular (Hip-Hop's Top Earners) · Lists (The Forbes 400)

Spit Happens! Genentech And 23andMe Team Up To Advance Genomic Testing In Clinical Trials

+ Comment Now + Follow Comments

Has genomic testing come of age?

I'm sure you've seen or heard the 23andMe commercials. It's the consumer-focused genomic testing service that, for $99, will analyze your DNA to provide an informational profile of hundreds of health conditions and trails. The key word here is "informational". These data are interesting, yet the exact clinical significance, in many instances is uncertain. It's the combination of this information with the combined wisdom of the medical community that can empower 23andMe. However, 23andMe is now teaming up with Genentech to profile a broad group of patients exposed to the Genentech drug Avastin (bevacizumab). The new consumer campaign is all about spit–and leverages the "playful ease' of getting a saliva sample that is rich with data–and dollars.

[TED Talk 2009 23andMe co-founder]

PHILANTHROPY 50

Comment (0)

February 10, 2013

No. 5: Sergey Brin and Anne Wojcicki

Kevork Djansezian/Getty Images

Sergey Brin and Anne Wojcicki

Enlarge Image

By Maria Di Mento

Amount donated in 2012: approximately $222.9-million

Top beneficiary: Brin Wojcicki Foundation

Other notable gift: $32.8-million to the Michael J. Fox Foundation for Parkinson's Research

Background: Mr. Brin co-founded Google. Ms. Wojcicki is a biotechnology analyst and co-founded the genetic-testing company, 23andME.

Mr. Brin and Ms. Wojcicki, both 39, gave about $190.1-million to their Brin Wojcicki Foundation, which supports a variety of causes. Last year the foundation awarded grants to Ashoka, which brings together social entrepreneurs to work on education, environment, women's issues, and many other causes; the Human Rights Foundation; and Tipping Point Community, a nonprofit that seeks to eliminate poverty in Northern California.

Your DNA suggests that if you take a blood thinner called warfarin (Coumadin®) you may require a lower dose.

Your DNA indicates that you are sensitive to the most common strain of norovirus.

You are a carrier for one or more inherited conditions.

Direct to consumer genetic testing

SHOW RESULTS FOR Chris Brown

SEE NEW AND RECENTLY UPDATED REPORTS »

Health Risks (120)

ELEVATED RISKS	YOUR RISK	AVERAGE RISK
Psoriasis	43.3%	11.4%
Chronic Kidney Disease	4.2%	3.4%
Scleroderma (Limited Cutaneous Type)	0.08%	0.07%

more »

DECREASED RISKS	YOUR RISK	AVERAGE RISK
Prostate Cancer ♂	8.2%	17.8%
Alzheimer's Disease	4.3%	7.2%

more »

See all 120 risk reports...

Inherited Conditions (50)

REPORT	RESULT
Hemochromatosis (HFE-related)	Variant Present
Familial Dysautonomia	Variant Absent
Canavan Disease	Variant Absent
Familial Hyperinsulinism (ABCC8-related)	Variant Absent
Primary Hyperoxaluria Type 2 (PH2)	Variant Absent
Rhizomelic Chondrodysplasia Punctata Type 1 (RCDP1)	Variant Absent
Torsion Dystonia	Variant Absent
Autosomal Recessive Polycystic Kidney Disease	Variant Absent

See all 50 carrier status...

Traits (60)

REPORT	RESULT
Alcohol Flush Reaction	Does Not Flush
Bitter Taste Perception	Can Taste
Earwax Type	Wet
Eye Color	Likely Blue
Hair Curl	Straighter Hair on Average

See all 60 traits...

Drug Response (24)

REPORT	RESULT
Warfarin (Coumadin®) Sensitivity	Increased
Clopidogrel (Plavix®) Efficacy	Reduced
Fluorouracil Toxicity	Typical
Sulfonylurea Drug Clearance (Type 2 Diabetes Treatment)	Typical
Alcohol Consumption, Smoking and Risk of Esophageal Cancer	Typical

See all 24 drug response...

Genome wide association studies

- E.g. for Multiple Sclerosis, Heart Disease, Depression, etc

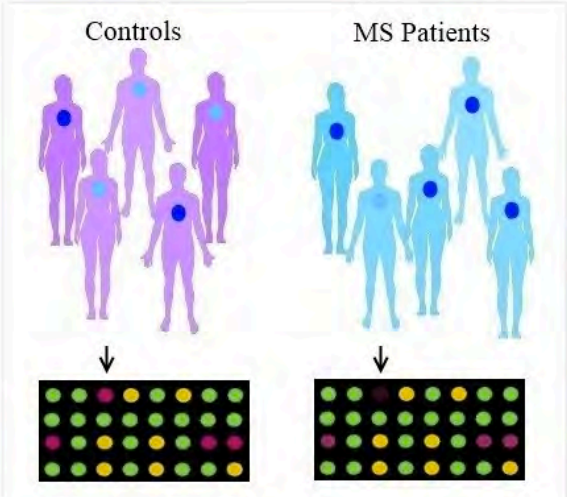
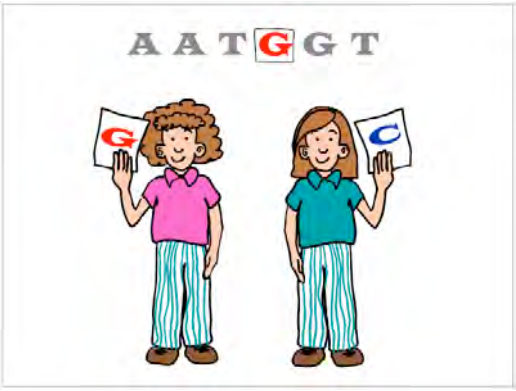


Fig. 2. Genome-wide association studies (GWAS) and MS. GWAS scan hundreds of thousands of points along the genome and identify single-nucleotide differences between MS cases and controls. In this image, one variant of a gene (represented by the light blue dot) is more prevalent among control subjects, whereas another variant of the same gene (represented by the dark blue dot) is more prevalent among people who have MS.



Single nucleotide polymorphisms (SNP)

Neanderthal?

This lab estimates your genome-wide percentage of Neanderthal ancestry

Got Neanderthal DNA?

An estimated 2.9% of your DNA is from Neanderthals.

Chris Brown (you) — 2.9% — 81st percentile

Average European user — 2.7%

MODERN HUMANS

Higher brow
Narrower shoulders
Slightly taller

NEANDERTHALS

Heavy eyebrow ridge
Long, low, bigger skull
Prominent nose with developed nasal chambers for cold-air protection

Friends & Family

You are ranked 1st among your friends. Invite more friends

Chris Brown (you) — 2.9% — 81st percentile among European users

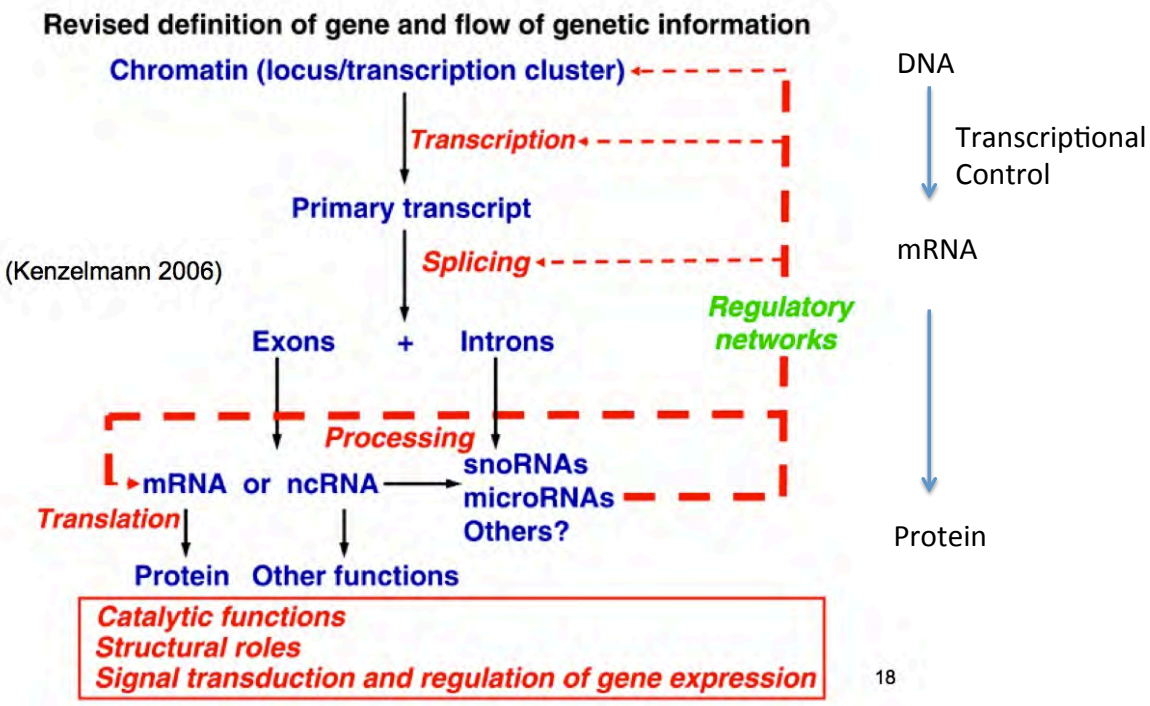
So what, I'm a caveman?

Actually yes, but that has little to do with the percentage of Neanderthal DNA in your genome. Our perception of Neanderthals as big oafs is clouded by our own notion of superiority and pop culture caricatures. How we are different and why modern humans survived and Neanderthals didn't is still mostly a mystery.

Neanderthal and proud?

Whatever your Neanderthal

Revision – Lecture 3/Slide18



Bioinformatics Research in Chris Brown group

Aim: Discovery of functional regulatory elements in (m)RNAs.

Experimental testing by ourselves or collaborators (NZ, USA, Swiss)

A. Databases of mRNAs and and mRNA regulatory elements

Webbased: Transterm, CRISPRTarget, CRISPRDetect

B. Methods for discovery of regulatory elements

Refinement or adaptation of existing methods for motif discovery, MEME, FIRE, etc . Results integration with SVM.

New algorithms – e.g. MLOGD, CDS-plotcon

Expression studies – microarrays and NGS

C. Workflows and Visualisation of elements on genomes

Galaxy browser for mRNA analysis (bioanalysis.otago.ac.nz)

Viral genomes (VirDB) using GBrowse (prototype: HBVRegDB)

Possible Summer project 2015 – Viral Genomic database, Virus genome analysid

D. Gene expression in plants

Effect of a bacterium on leaves – Model system (Plant and Food Auckland)

Beneficial fungus (endophyte) on Maize roots. (Lincoln)

People past and present - Brown lab

Dream to turn sea urchins into lucrative export industry

Home » News » Dunedin

By John Gibb on Sat, 27 Aug 2011

University of Otago | News: Dunedin | Graduation

Share 0 Tweet 0 ShareThis

Daniel Garama has realised his dream of becoming a biochemist and his research on sea urchins could help develop a lucrative aquaculture export industry.

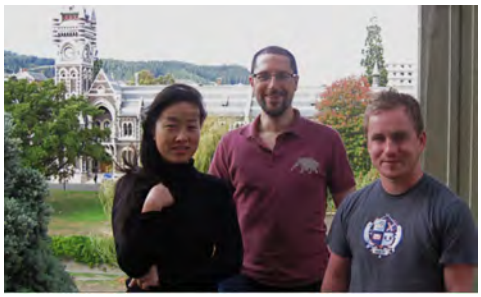
Mr Garama (30) will today gain a doctorate from the University of Otago, as one of about 350 graduands in all disciplines who will graduate in person in a 3pm ceremony at the Regent Theatre.

"Obviously, I'm proud," he said yesterday.

He felt "a sense of fulfilment" to have finally reached his goal, after many years of hard work.

Biochemist Daniel Garama, who graduates from the University of Otago with a PhD today, looks at sea urchin shells. Photo by Craig Baxter.

Dr Dan Garama (Cancer, Kina)



Some of Brown Group: Dr Sylvia Chen (now in Netherlands –Regulatory elements), Stewart Stevens (PhD student- Gene expression, MEd) Josh Gagnon (Software developer- lots)

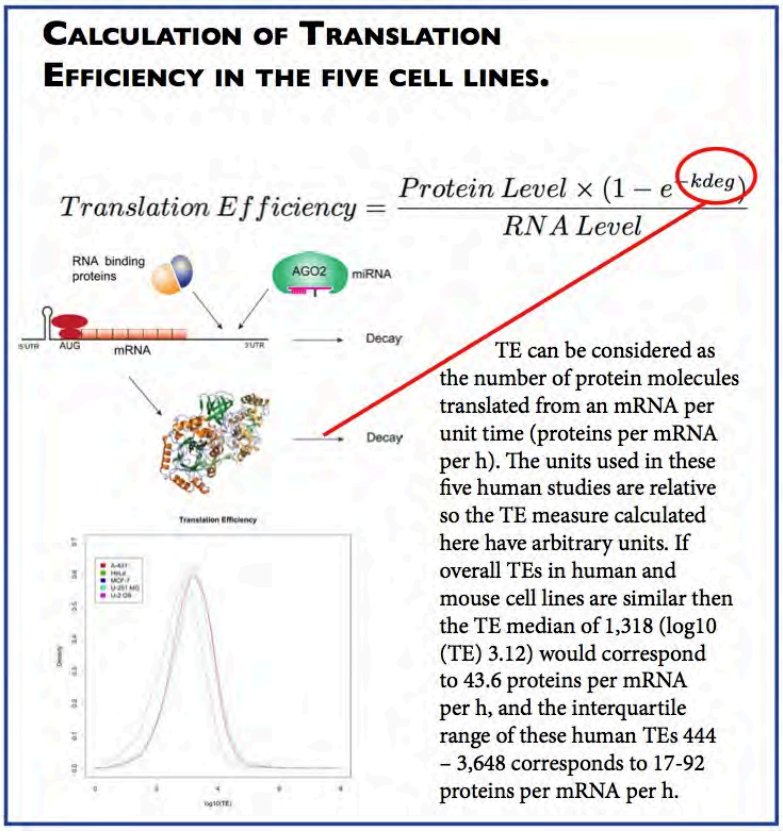
2013

Andrew Sarman (MSc Genetics) – Breast cancer

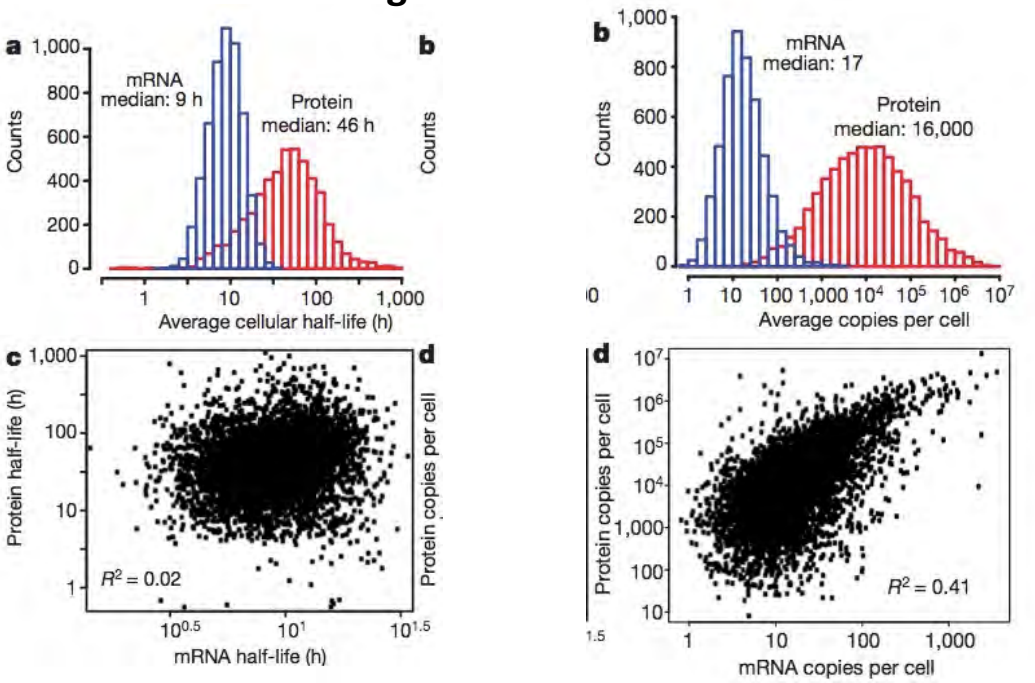
Scout Liu (Hons, Mol Biotech) – Cancer- MYC transcription factor

Ambarish Biswas (PhD Biochem) – Regulatory elements

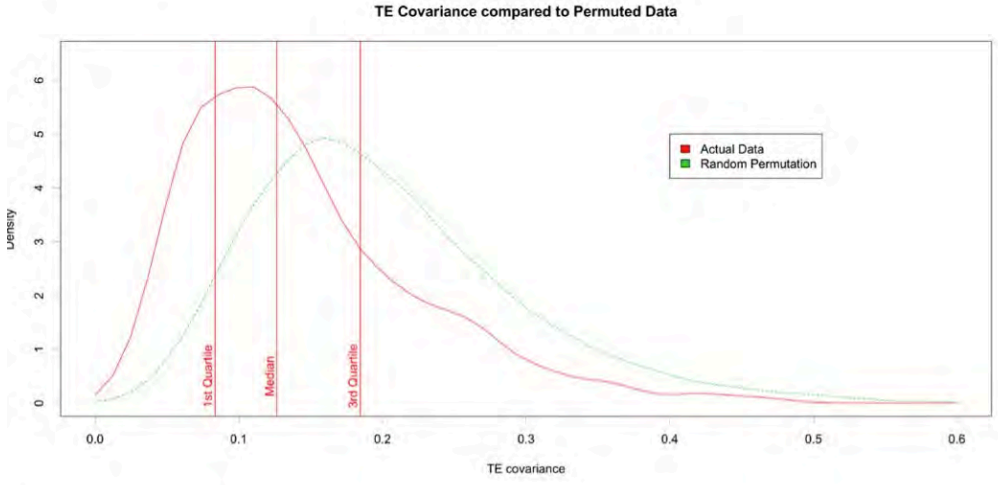
Gareth Gillard (MSc Biochem) – Gene expression, Genomics, RNA-seq.



‘Good’ correlation between protein and mRNA copies However, no correlation between stabilities! – complex regulation

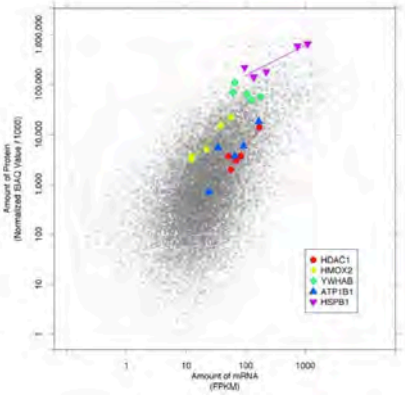


Translation Efficiency shows greater consistency than expected by chance within Genes across Cell Lines.



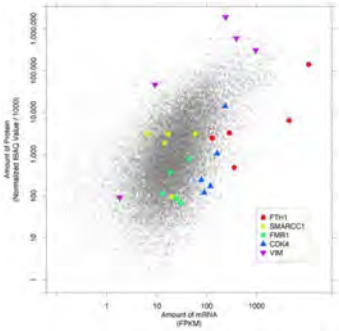
We did a random permutation of TE values and compared the resulting covariance with the actual covariance observed within each gene.

GENES WITH A LOW COEFFICIENT OF VARIATION (CV) IN TE VALUE.



Some examples of well-studied genes are shown here. They tend to either have similar expression values at both the protein and mRNA level (e.g. YWHAB) or have linear relationships between these values (e.g. HMOX2, HSPB1).

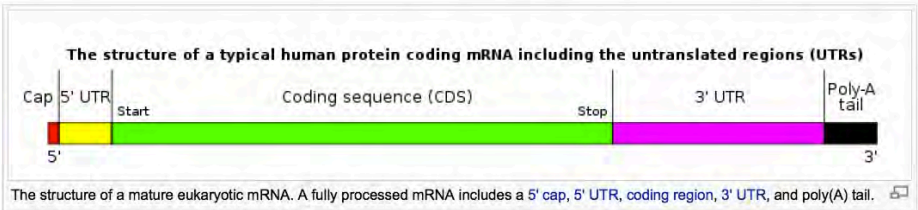
GENES WITH A HIGH COEFFICIENT OF VARIATION (CV) IN TE VALUE.



Examples shown here include FTH1, which has well studied translational control mechanisms. Genes such as SMARCC1 have large variation in the amount of mRNA with smaller variation in the amount of protein. Conversely genes such as CDK4 have widely ranging amounts of protein but little difference in the amounts of mRNA. Other genes such as VIM vary in both protein and mRNA amounts.

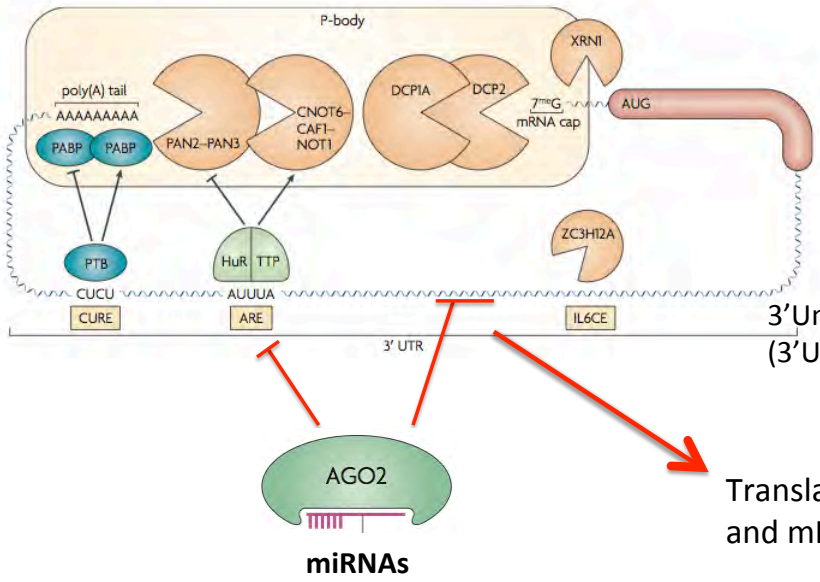
Human mRNA structure

Structure [edit]



The structure of a mature eukaryotic mRNA. A fully processed mRNA includes a 5' cap, 5' UTR, coding region, 3' UTR, and poly(A) tail.

Interactions with mRNA and its decay
-Post-transcriptional control may be through the 3'UTR



Translated region

3'Untranslated region (3'UTR)

Translational repression and mRNA decay

miRNAs

Paul Anderson
Nature Reviews Immunology 2010

Review

Cell PRESS

Feature Review

De novo prediction of structured RNAs from genomic sequences

Jan Gorodkin[1], Ivo L. Hofacker[2], Elfar Torarinsson[1], Zizhen Yao[3], Jakob H. Havgaard[1] and Walter L. Ruzzo[3,4]

[1] Section for Genetics and Bioinformatics, IBHV and Center for Applied Bioinformatics, University of Copenhagen, Grønnegårdsvej 3, DK-1870 Frederiksberg C, Denmark
[2] Institut für theoretische Chemie, University of Vienna, Währingerstr. 17, A-1090 Vienna, Austria
[3] Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N. PO Box 19024, Seattle, Washington 98109, USA
[4] Departments of Computer Science and Engineering and Genome Sciences, University of Washington, Seattle, Washington 98195, USA

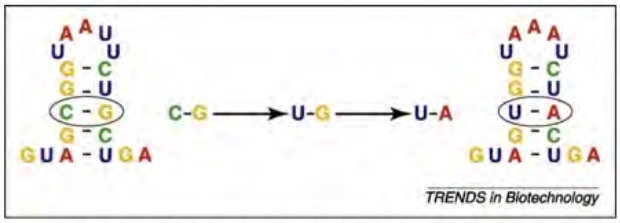
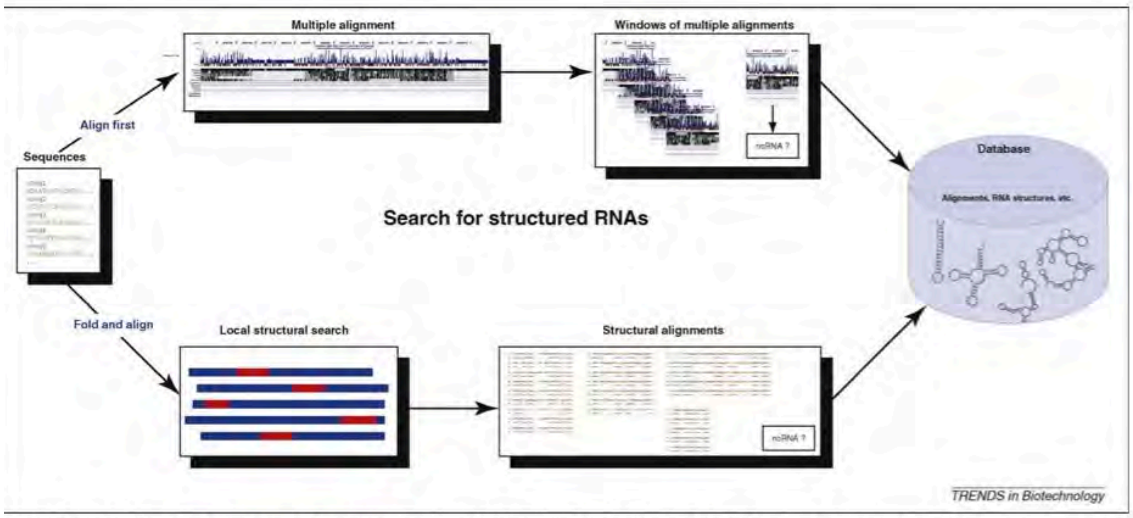


Figure 1. Compensating base changes. Changes in base pairing might preserve structure, but not the primary sequence. In addition to the usual Watson–Crick base pairs, less stable G–U pairs (sometimes called "wobble pairs") are often seen in RNAs, and are evolutionarily important because they allow single base substitutions that are not structurally disruptive. This might allow sequences to accumulate substitutions much more rapidly than would be the case if both nucleotides in a base pair needed to be changed more or less simultaneously. Adapted with permission from Ref. [80].



igure 2. Strategies for ncRNA screening. The upper path illustrates RNA structure prediction using existing sequence alignments that are divided into overlapping indows (an align first strategy). By contrast, in the lower path, labeled "fold and align," sequence and structure alignments are performed directly from unaligned equence data (a joint strategy), which searches simultaneously for conserved structure and sequence, and results in structural alignments. To date, alternative fold-first trategies have not been applied to genome-scale screening.

Little overlap between algorithms

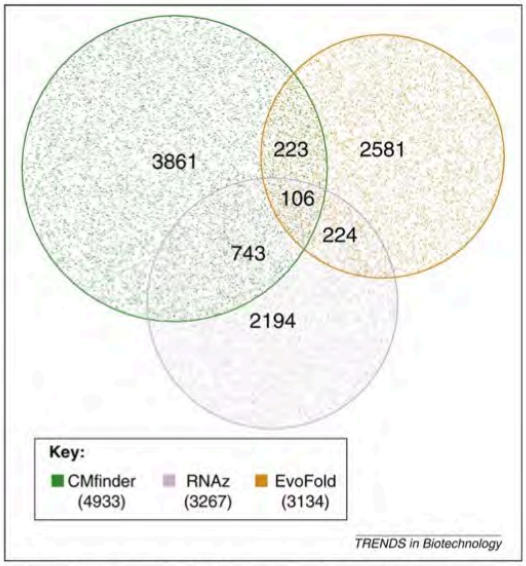
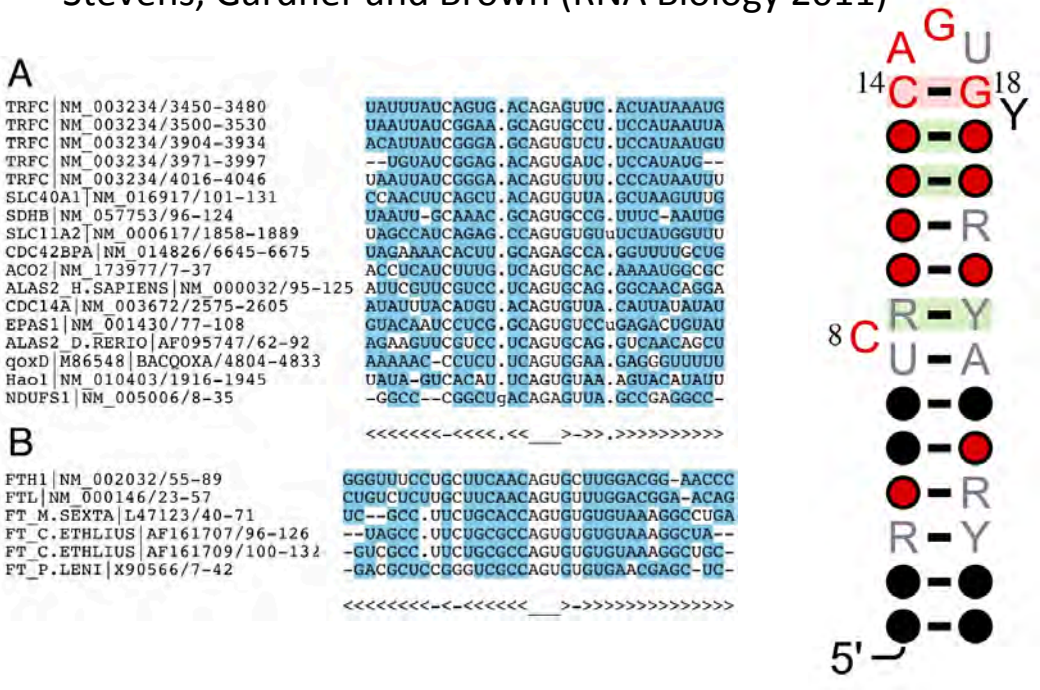


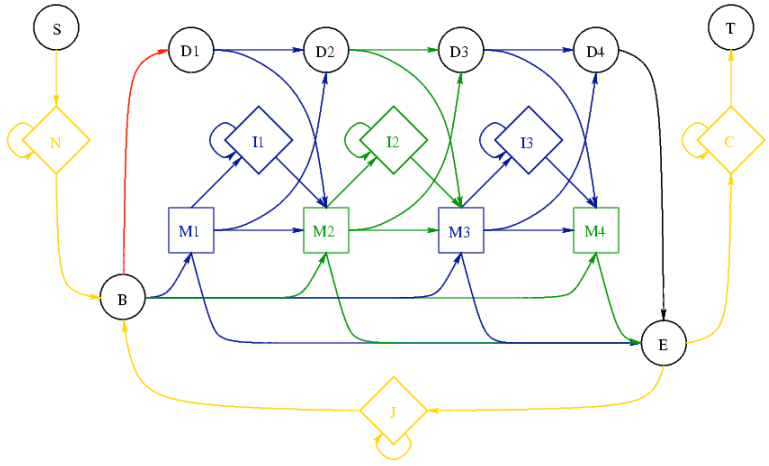
Figure 6. Comparison of ENCODE scans. The Venn diagram compares RNA elements in the ENCODE regions predicted by the three screening methods discussed here, CMfinder, RNAz, and EvoFold. Only predictions from input data that are common to all three studies are tallied, that is, repeat, exon and PhastCons regions are excluded [64]. As illustrated here, of the 4933 candidates reported by CMfinder, 3861 were reported by neither of the other methods, while 106 candidates were reported by all three. Adapted with permission from [Ref. 64].

More complex models of regulatory elements
e.g. Iron responsive Element
Covariance Model from experimentally verified examples (Rfam)

Stevens, Gardner and Brown (RNA Biology 2011)

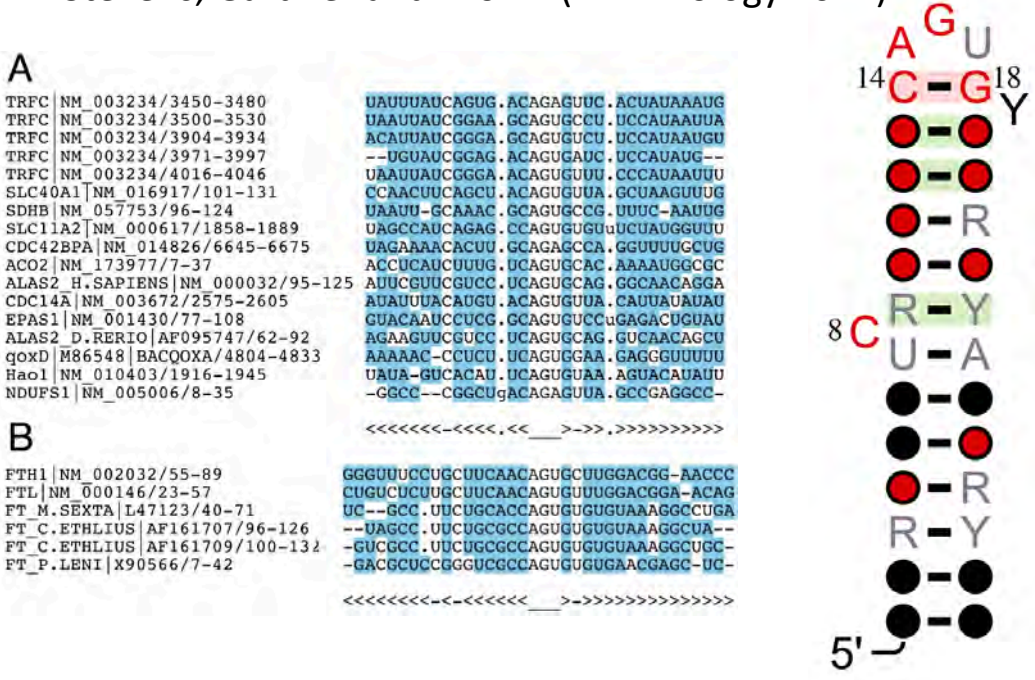


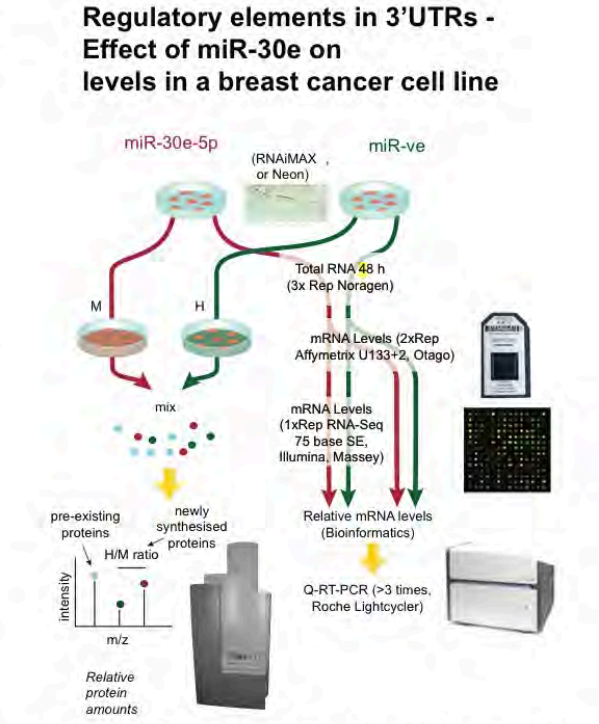
Covariance model – an elaboration of a HMM
Additional terms for covariation



More complex models of regulatory elements
e.g. Iron responsive Element
Covariance Model from experimentally verified examples (Rfam)

Stevens, Gardner and Brown (RNA Biology 2011)

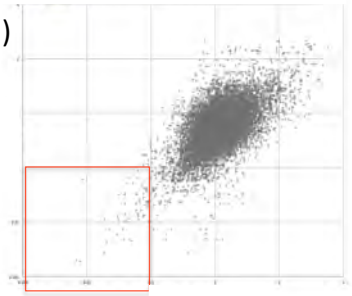




The breast cancer cell line hs.578-T is a triple negative cell line, representing a type of breast cancer for which there is limited trageted treatment. Specific proteins and miRNAs are up and downregulated and we are investigating there role in carcinogenesis. Testing the effect of increasing the levels of miR-30e will identify targets regulated by mRNA stability or by inhibition of translation.

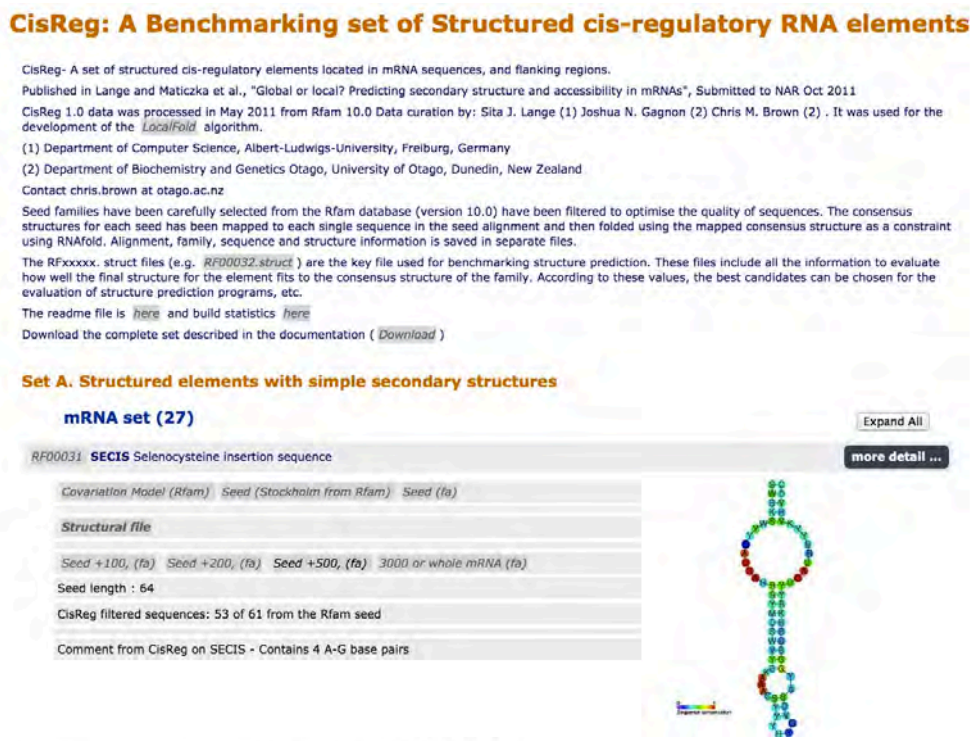
Changes in mRNA levels by RNA-Seq

- ~30 million SE, 75 base Illumina reads (Massey).
- 304 ≤ 0.5 down, 100 ≥ 2 up of 13,792 transcripts with > 50 reads in control.
- Ratios good correlation with array and Q-PCR
- Examples of mRNAs with changes in levels:



Gene	miR-ve (reads)	miR-30e (norm reads)	Ratio	Ratio from Array	Prediction
ZDHHC20	418	60.0	0.14	0.21	1
IDH1	2848	464.7	0.16	0.22	4
SEC23A	6902	1247.5	0.18	0.3	5
KDELC2	3552	733.5	0.21	0.31	3
SPTLC3	421	87.8	0.21	0.43	4
TWF1	1972	414.4	0.21	0.4	5
GM2A	675	145.6	0.22	0.4	3
RARG	868	187.4	0.22	NS	4
EXTL2	686	157.4	0.23	0.34	5
CD99	3701	905.9	0.24	0.37	2
NT5E	6486	1592.3	0.25	0.32	6
MAP2K6	64	17.1	0.27	NS	1
AP3S1	961	258.1	0.27	0.49	3
PEG10	951	260.2	0.27	0.38	3

Predictions - Votes 0-6
MiR-30e target prediction was from
TargetScan (3% of genes predicted positive)
Miranda (15%)
mirTarget2 (3%)
Pictar (4%)
Pita (29%)
RNAHybrid (97%),



CisReg: A Benchmarking set of Structured cis-regulatory RNA elements

CisReg- A set of structured cis-regulatory elements located in mRNA sequences, and flanking regions.

Published in Lange and Maticzka et al., "Global or local? Predicting secondary structure and accessibility in mRNAs", Submitted to NAR Oct 2011

CisReg 1.0 data was processed in May 2011 from Rfam 10.0 Data curation by: Sita J. Lange (1) Joshua N. Gagnon (2) Chris M. Brown (2) . It was used for the development of the *LocalFold* algorithm.

(1) Department of Computer Science, Albert-Ludwigs-University, Freiburg, Germany

(2) Department of Biochemistry and Genetics Otago, University of Otago, Dunedin, New Zealand

Contact chris.brown at otago.ac.nz

Seed families have been carefully selected from the Rfam database (version 10.0) have been filtered to optimise the quality of sequences. The consensus structures for each seed has been mapped to each single sequence in the seed alignment and then folded using the mapped consensus structure as a constraint using RNAfold. Alignment, family, sequence and structure information is saved in separate files.

The RFxxxxx. struct files (e.g. *RF00032.struct*) are the key file used for benchmarking structure prediction. These files include all the information to evaluate how well the final structure for the element fits to the consensus structure of the family. According to these values, the best candidates can be chosen for the evaluation of structure prediction programs, etc.

The readme file is *here* and build statistics *here*

Download the complete set described in the documentation (*Download*)

Set A. Structured elements with simple secondary structures

mRNA set (27)

Expand All

RF00031 **SECIS** Selenocysteine insertion sequence — more detail ...

Covariation Model (Rfam) *Seed (Stockholm from Rfam)* *Seed (fa)*

Structural file

Seed +100, (fa) *Seed +200, (fa)* ***Seed +500, (fa)*** *3000 or whole mRNA (fa)*

Seed length : 64

CisReg filtered sequences: 53 of 61 from the Rfam seed

Comment from CisReg on SECIS - Contains 4 A-G base pairs

Published online 28 February 201.

Global or local? Predicting secondary structure and accessibility in mRNAs

Sita J. Lange[1], Daniel Maticzka[1], Mathias Möhl[1], Joshua N. Gagnon[2], Chris M. Brown[2] and Rolf Backofen[1,*]

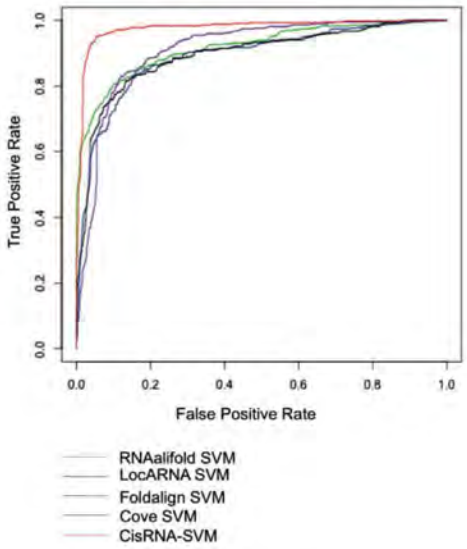
[1]Department of Computer Science and Centre for Biological Signalling Studies (BIOSS), Albert-Ludwigs-Universität Freiburg, Germany and [2]Department of Biochemistry and Genetics Otago, University of Otago, P.O. Box 56, 710 Cumberland St, Dunedin 9054, New Zealand

Computational identification of new structured *cis*-regulatory elements in the 3′-untranslated region of human protein coding genes

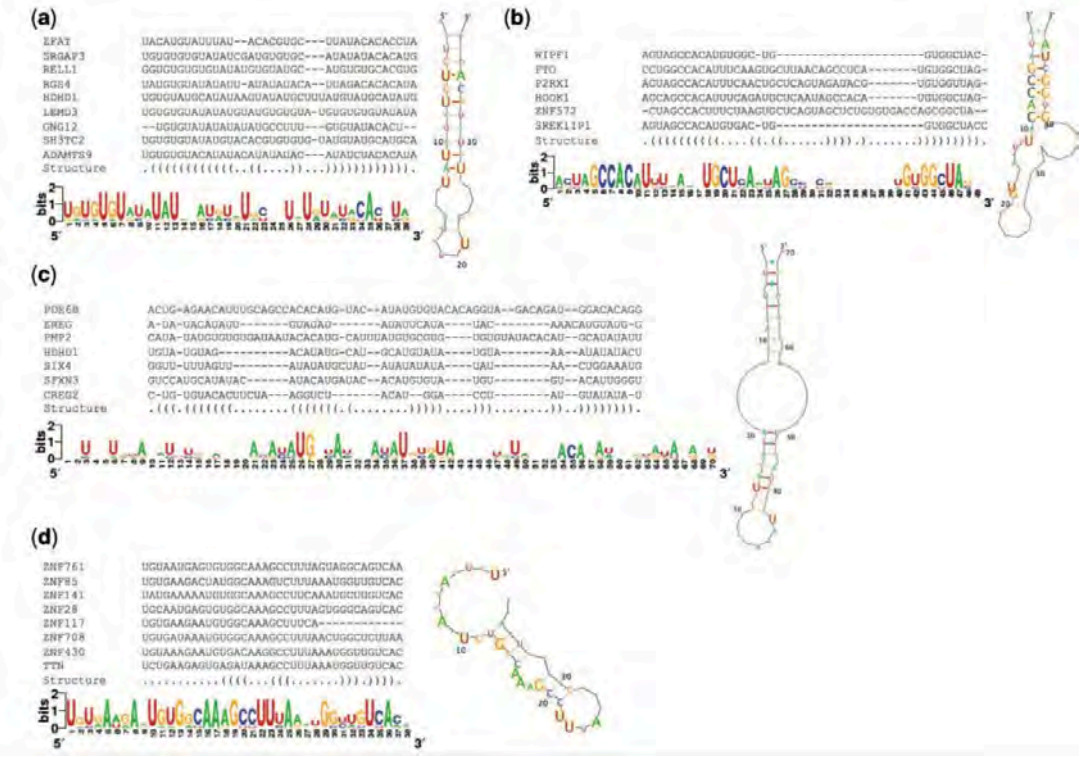
Xiaowei Sylvia Chen* and Chris M. Brown

Department of Biochemistry and Genetics Otago, University of Otago, Dunedin 9054, New Zealand

Received April 19, 2012; Revised June 15, 2012; Accepted June 20, 2012



Predicted novel elements



NeSI and high RAM computers

STAT 435

STAT435 Data Analysis for Bioinformatics

First Semester, 20 points

Resources

The analysis of large data sets is becoming increasingly important in many areas. The techniques covered in this course will be applicable to a wide range of data types, including non-biological data. Exposure to other disciplines (in this case biomedical science) is a must for any applied statistician. Interacting with students from other fields is simulating, and will help you appreciate your statistical skills.

Paper details

- Overview of genetics and molecular biology.
- Introduction to genetic, genomic, and proteomic technologies.
- Methods for the statistical analysis of large data sets.
- Application of standard statistical methods.
- Introduction to new purpose-built methods.
- Incorporation of biological information into the statistical analysis process.

Potential students

This course is open to 4th year students from the biological and medical sciences, mathematics and statistics, and computer science. As long as you have skills in one of these areas, any remaining gaps will be filled in during the course. Experience with R will certainly help.

Bioinformatics Research in Chris Brown group

Aim: Discovery of functional regulatory elements in (m)RNAs.

Experimental testing by ourselves or collaborators (NZ, USA, Swiss)

A. Databases of mRNAs and and mRNA regulatory elements

Webbased: Transterm, CRISPRTarget, CRISPRDetect

B. Methods for discovery of regulatory elements

Refinement or adaptation of existing methods for motif discovery, MEME, FIRE, etc . Results integration with SVM.

New algorithms – e.g. MLOGD, CDS-plotcon

Expression studies – microarrays and NGS

C. Workflows and Visualisation of elements on genomes

Galaxy browser for mRNA analysis (bioanalysis.otago.ac.nz)

Viral genomes (VirDB) using GBrowse (prototype: HBVRegDB)

Possible Summer project 2015 – Viral Genomic database, Virus genome analysid

D. Gene expression in plants

Effect of a bacterium on leaves – Model system (Plant and Food Auckland)

Beneficial fungus (endophyte) on Maize roots. (Lincoln)