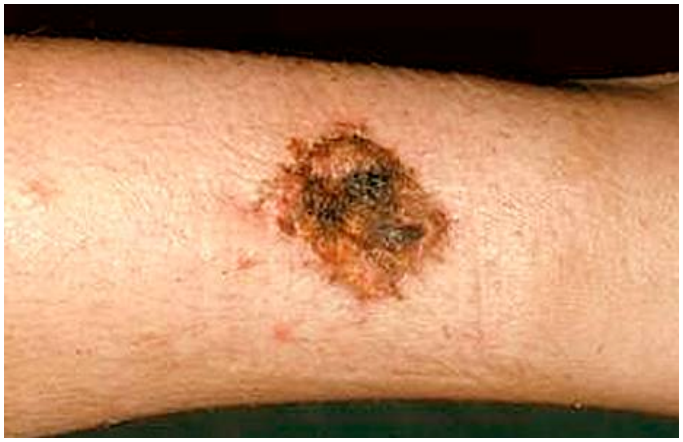
MicroRNA Analysis by Hypothesis Finding Techniques

Andrei Doncescu
LAAS CNRS/ Toulouse France
NII Tokyo Japan

Katsumi Inoue
NII Tokyo
Japan



Objective :

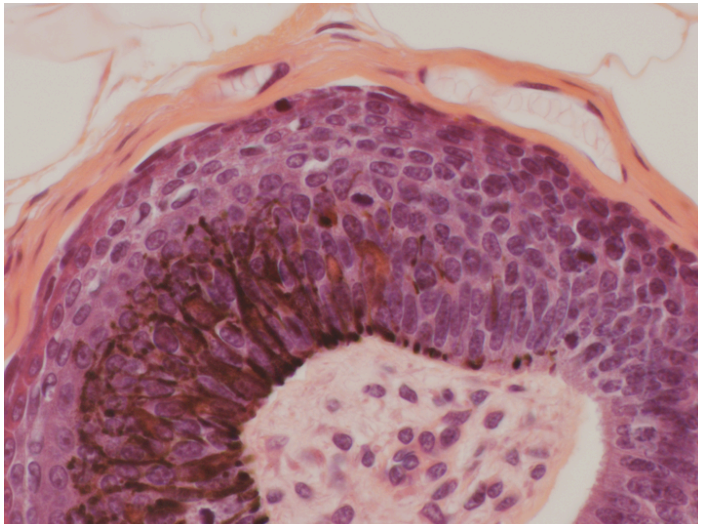
- **Find a Plasma/Serum MicroRNA Signature (CNF form)**

  able to categorize populations :

  – Without melanoma

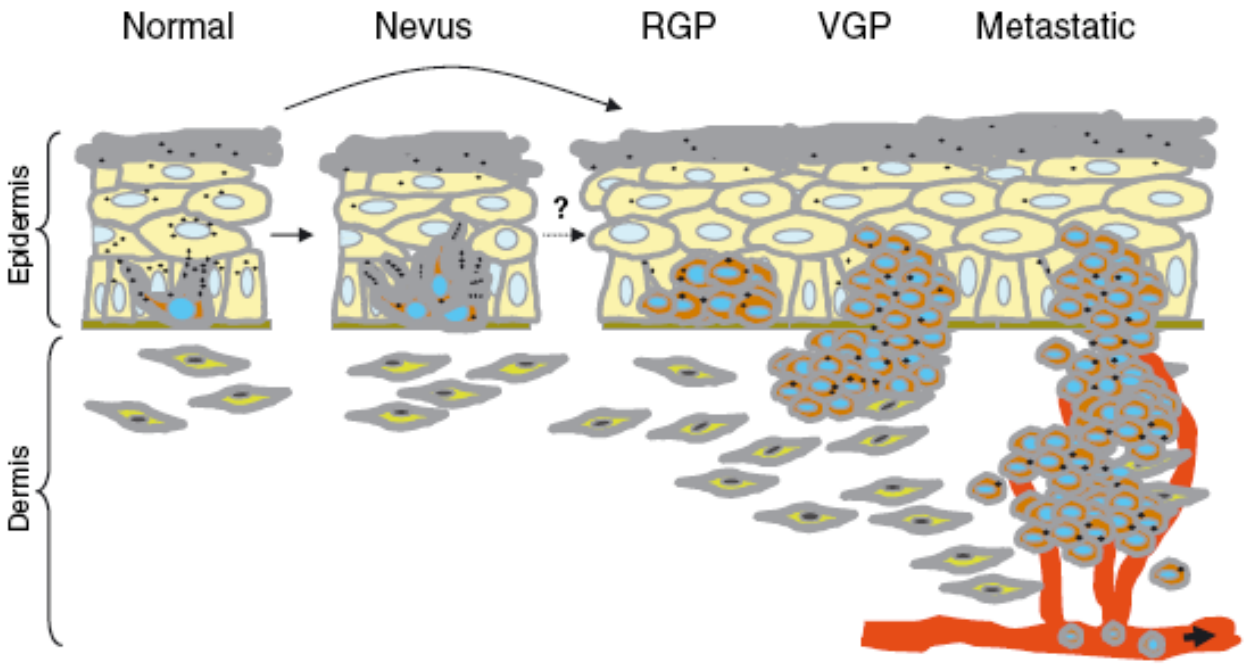
  – With Melanoma

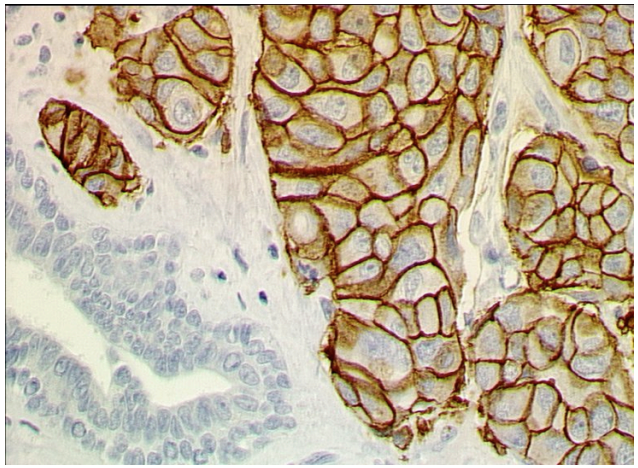
    - Fast development of the cancer : metastasis



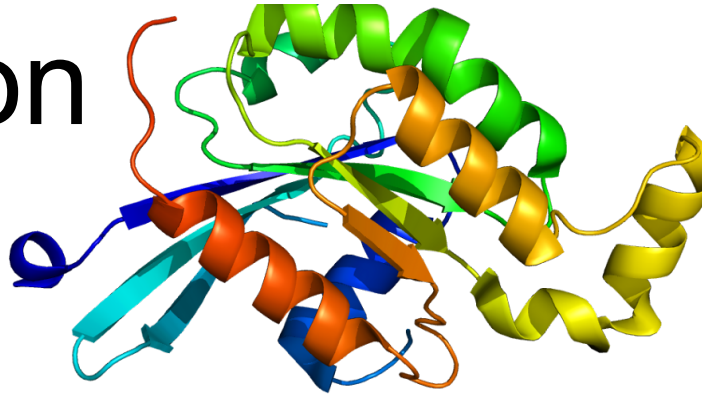
Melanoma

- 7231 new cases were diagnosed in 2000 in France
- The treatment of melanoma in the metastatic stage is disappointing
- Only 10% of the patients respond to the treatment but often only partially and for short time.





Gene expression

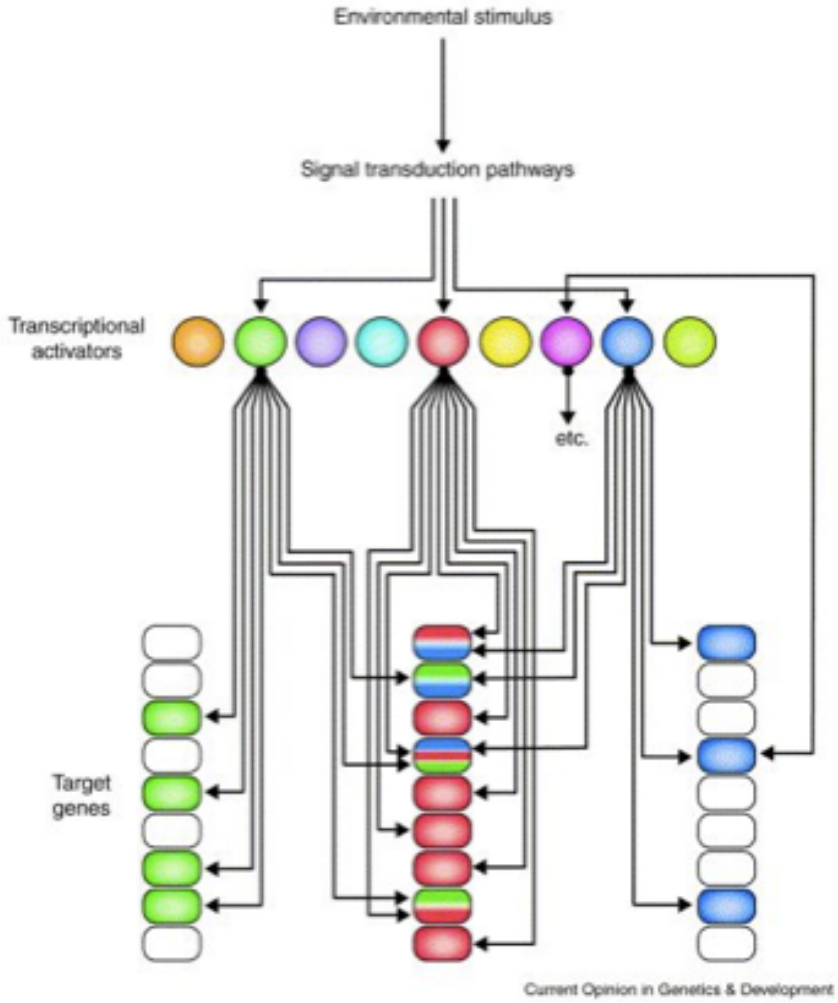


- Distinct cellular identities are due to gene expression
  (= transcription & translation of gene).

- Transcription generates 3 kinds of RNA : **mRNA**, **tRNA**, **rRNA.**

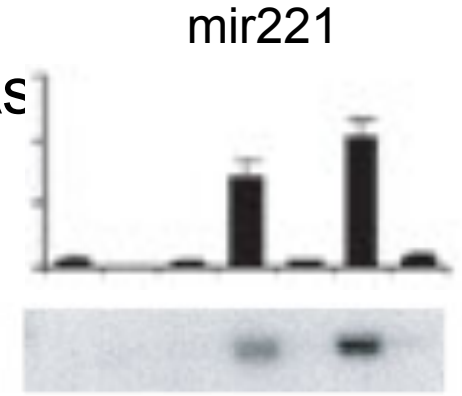
- Whether a gene is transcribed is often determined by the presence/ absence of other gene products (proteins)
- genes interact in complex networks: gene A switches on B, which turns off C which **upregulates** (increases) A, ..

- Perturbations to single gene can lead to changes in expression of many genes.



microRNA (1993)

- A Genome has protein-coding genes
- It also has genes that code for RNA
  - “transfer RNA” that is used in translation is coded by genes
  - “ribosomal RNA” that forms part of the structure of the ribosome, is also coded by genes
- microRNAs are a family of small RNAs
  - A genome has genes that code for microRNAs i.e., the result of transcription is microRNA



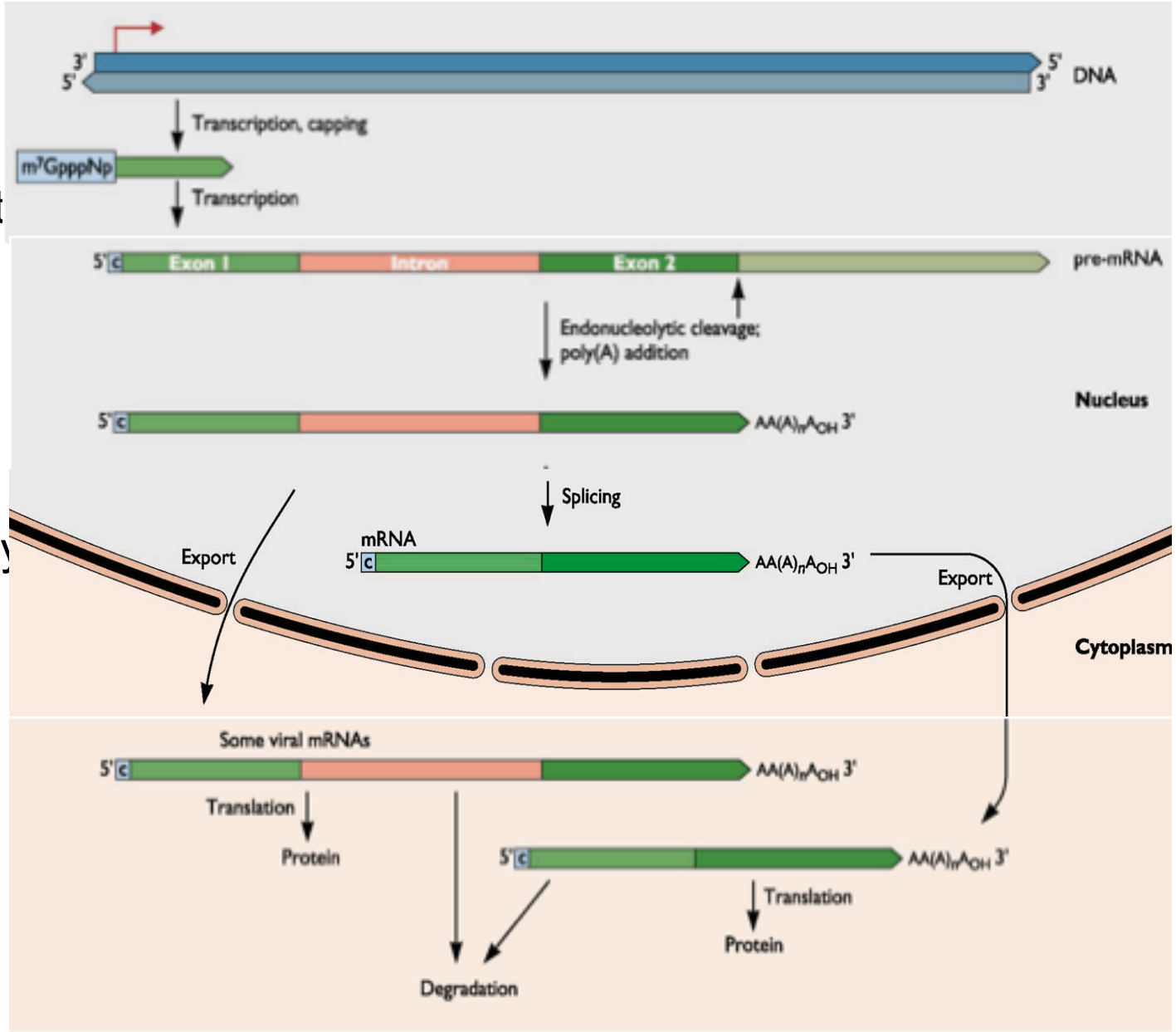
microRNA

The Vast majority of microRNAs regulate other genes by binding complementary sequences in the target gene

21-22 nucleotide non-coding RNA

Perfect complementarity of binding leads to mRNA degradation of the target gene

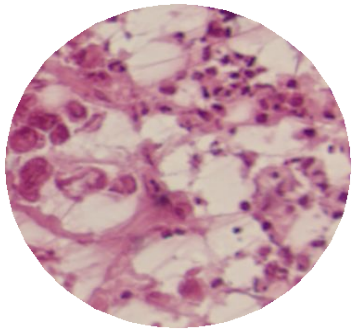
Imperfect pairing inhibits translation of mRNA to a protein



Cancer Analysis

Core premise:

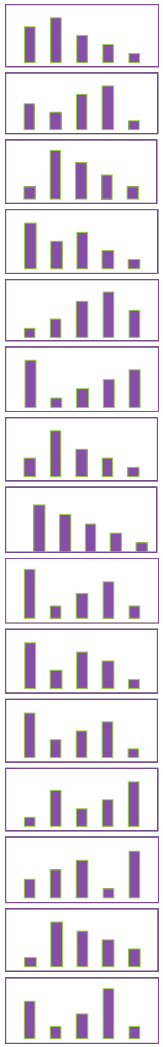
- Different tissue types have distinct mRNA profiles



Uncertain specimen RNA profile generated

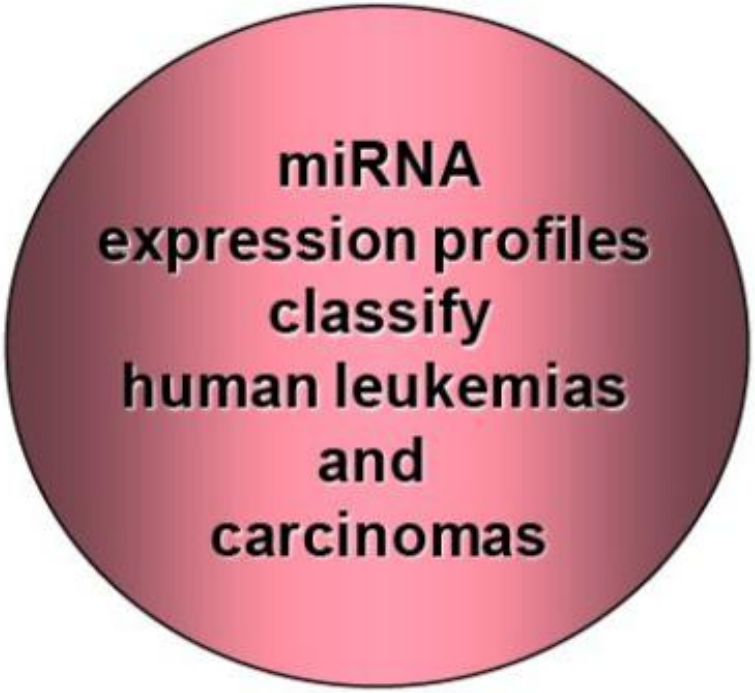
>2,000 genes

RNA profile compared with known tissue profiles



Similarity Scores Generated	
Colorectal	88.2
Pancreas	4.4
Non-small Cell Lung	2.3
Breast	2.1
Gastric	1.2
Kidney	0.6
Hepatocellular	0.3
Ovarian	0.3
Soft Tissue Sarcoma	0.1
Non-Hodgkin's Lymphoma	0.1
Thyroid	0.1
Prostate	0.1
Melanoma	0.1
Bladder	0.1
Testicular Germ Cell	0.0

ALTERATIONS OF MICRORNAS ARE FOUND IN EVERY TYPE OF HUMAN CANCER



p27 or the cyclines D1, D3 in Melanoma

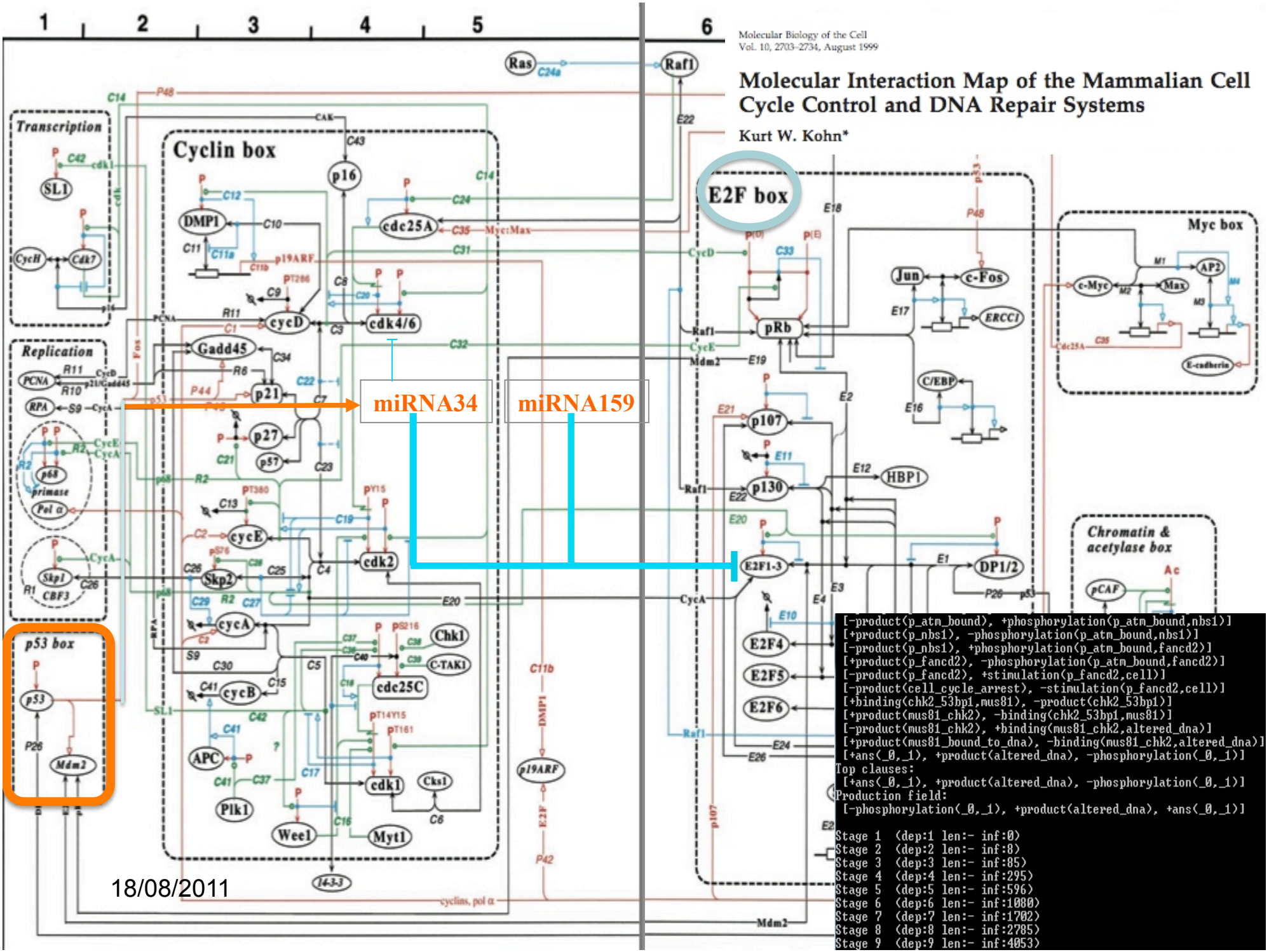
(Calin et al, PNAS 2002; Lu et al, Nature, 2005; Volinia & Calin et al, PNAS 2006; Landgraf et al, Cell 2007)

miRNA Oncogenes or Tumor Suppressor Genes (Croce Nat Rev Genet. 2009 Oct;10(10):704-14.)

Table 1 | **MicroRNAs that function as oncogenes or tumour suppressor genes in human cancers**

MicroRNA	Dysregulation	Function	Validated targets	Oncogene (ONC) or tumour suppressor (TS)	Refs
miR-15a and *miR-16-1*	Loss in CLL, prostate cancer and multiple myeloma	Induces apoptosis and inhibits tumorigenesis	BCL2, WT1 RAB9B and MAGE83	TS	15,20,23, 30,52,69
let-7 (*a*, *b*, *c*, *d*, *e*, *f*, *g* and *i*)	Loss in lung and breast cancer and in various solid and haematopoietic malignancies	Induces apoptosis and inhibits tumorigenesis	RAS, MYC and HMGA2	TS	22,26, 42,70
miR-29 (*a*, *b* and *c*)	Loss in aggressive CLL, AML (11q23), MDS lung and breast cancers and cholangocarcinoma	Induces apoptosis and inhibits tumorigenicity. Reactivates silenced tumour suppressor genes	TCL1, MCL1 and DNMTs	TS	30,64, 71,72
miR-34	Loss in pancreatic, colon, breast and liver cancers	Induces apoptosis	CDK4, CDK6, cyclin E2, EZF3 and MET	TS	56–58
miR-145	Loss in breast cancer	Inhibits proliferation and induces apoptosis of breast cancer cells	ERG	TS	31
miR-221 and *miR-222*	Loss in erythroblastic leukaemia	Inhibits proliferation in erythroblasts	KIT	TS	30
miR-221 and *miR-222*	Overexpression in aggressive CLL, thyroid carcinoma and hepatocellular carcinoma	Promotes cell proliferation and inhibits apoptosis in various solid malignancies	p27, p57, PTEN and TIMP3	ONC	43,51,73
miR-155	Upregulated in aggressive CLL, Burkitt's lymphoma and lung, breast and colon cancers	Induces cell proliferation and leukaemia or lymphoma in mice	MAF and SHIP1	ONC	32–34, 36,37
miR-17–92 cluster	Upregulated in lymphomas and in breast, lung, colon, stomach and pancreatic cancers	Induces proliferation	E2F1, BIM and PTEN	ONC	19,34,35, 40,41
miR-21	Upregulated in glioblastomas, AML (11q23), aggressive CLL and breast, colon, pancreatic, lung, prostate, liver and stomach cancers	Inhibits apoptosis and increases tumorigenicity	PTEN, PDCD4, TPM1 and TIMP3	ONC	31,37–39, 44–50
miR-372 and *miR-373*	Upregulated in testicular tumours	Promotes tumorigenicity in cooperation with RAS	LATS2	ONC	74

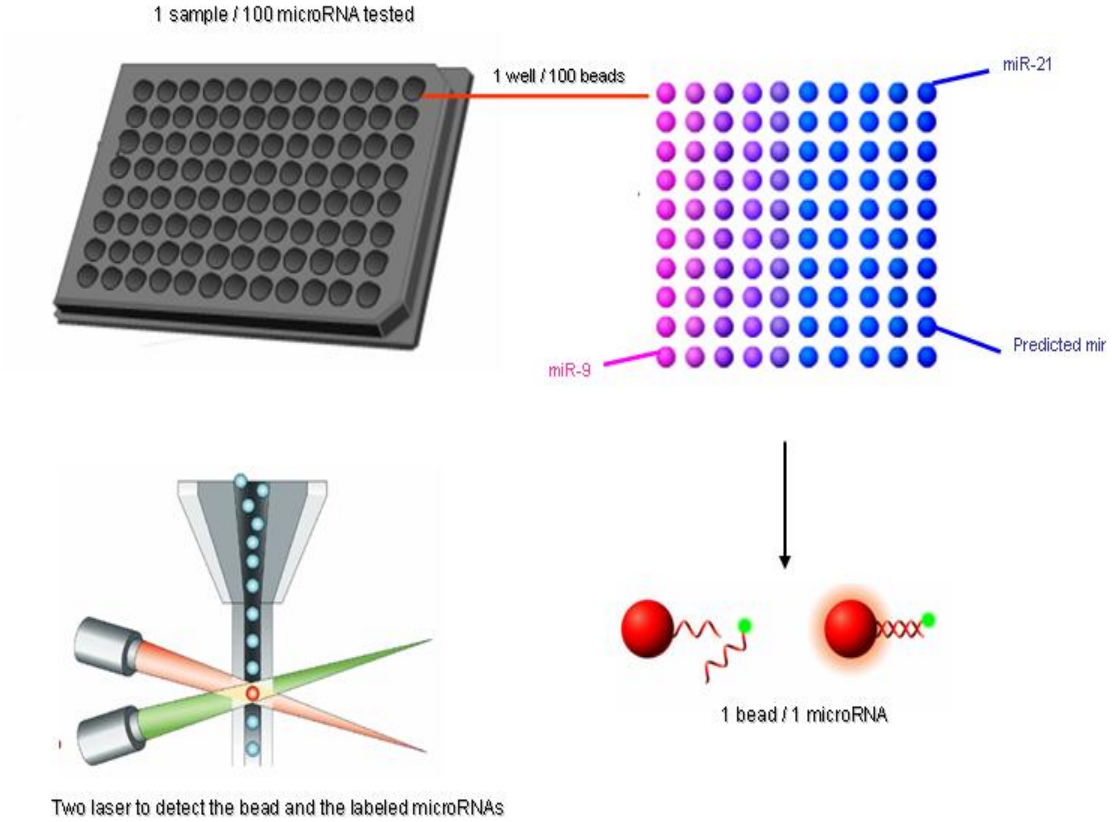
AML, acute myeloid leukaemia; BCL2, B cell leukaemia/lymphoma 2; BIM, Bcl2-interacting mediator of cell death; CLL, chronic lymphocytic leukaemia; DNMT, DNA methyltransferase; HMGA2, high mobility group AT-hook 2; LATS2, large tumour suppressor homologue 2; MCL1, myeloid cell leukaemia sequence 1; MDS, myelodysplastic syndrome; PDCD4, programmed cell death 4; PTEN, phosphatase and tensin homologue; SHIP1, SH2 domain-containing inositol-5′-phosphatase 1; TCL1, T cell lymphoma breakpoint 1; TIMP3, tissue inhibitor of metalloproteinases 3; TPM1, tropomyosin 1; WT1, Wilms tumour 1.

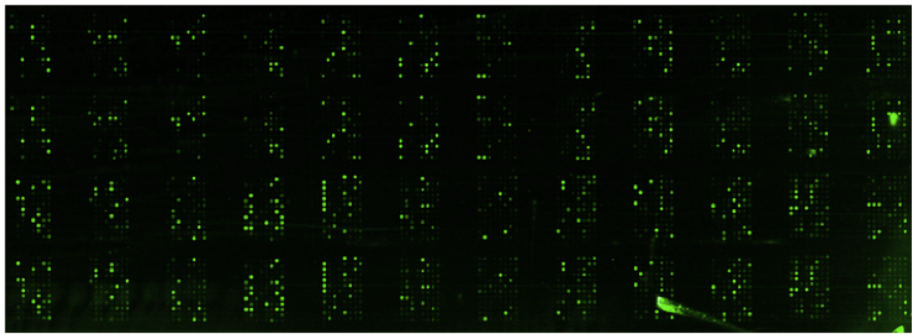

1
2
3
4
5
6
Molecular Biology of the Cell
Vol. 10, 2703–2734, August 1999
Molecular Interaction Map of the Mammalian Cell Cycle Control and DNA Repair Systems
Kurt W. Kohn*
Transcription
Cyclin box
E2F box
Myc box
Replication
p53 box
Chromatin & acetylase box
miRNA34
miRNA159
Ras
Raf1
p16
DMP1
cdc25A
p19ARF
cycD
cdk4/6
Gadd45
p21
p27
p57
cycE
cdk2
Skp2
cycA
Chk1
C-TAK1
cdc25C
cycB
APC
Plk1
Wee1
Myt1
cdk1
Cks1
14-3-3
pRb
p107
p130
E2F1-3
DP1/2
E2F4
E2F5
E2F6
HBP1
Jun
c-Fos
ERCC1
c-Myc
Max
AP2
E-cadherin
C/EBP
pCAF
p53
Mdm2
[-product(p_atm_bound), +phosphorylation(p_atm_bound,nbs1)]
[+product(p_nbs1), -phosphorylation(p_atm_bound,nbs1)]
[-product(p_nbs1), +phosphorylation(p_atm_bound,fancd2)]
[+product(p_fancd2), -phosphorylation(p_atm_bound,fancd2)]
[-product(p_fancd2), +stimulation(p_fancd2,cell)]
[-product(cell_cycle_arrest), -stimulation(p_fancd2,cell)]
[+binding(chk2_53bp1,mus81), -product(chk2_53bp1)]
[+product(mus81_chk2), -binding(chk2_53bp1,mus81)]
[-product(mus81_chk2), +binding(mus81_chk2,altered_dna)]
[+product(mus81_bound_to_dna), -binding(mus81_chk2,altered_dna)]
[+ans(_0,_1), +product(altered_dna), -phosphorylation(_0,_1)]
Top clauses:
[+ans(_0,_1), +product(altered_dna), -phosphorylation(_0,_1)]
Production field:
[-phosphorylation(_0,_1), +product(altered_dna), +ans(_0,_1)]
Stage 1 (dep:1 len:- inf:0)
Stage 2 (dep:2 len:- inf:8)
Stage 3 (dep:3 len:- inf:85)
Stage 4 (dep:4 len:- inf:295)
Stage 5 (dep:5 len:- inf:596)
Stage 6 (dep:6 len:- inf:1080)
Stage 7 (dep:7 len:- inf:1702)
Stage 8 (dep:8 len:- inf:2785)
Stage 9 (dep:9 len:- inf:4053)
18/08/2011

Profile Analysis of Plasma miRNA Healthy Patients/Metastatic Melanoma

➔Analyse of Plasma miRNAs on oligonucletides modified (LNA):

➔5 healthy subjects (EFS) versus 15 patients with melanoma





87 to 98 miRNA for 10 ml to 50 ml of plasma. Mitchell PS. et al, *PNAS, 2008:* Chen X. et al, *Cell Research, 2008*

	melanoma	normal
DOWN		
hsa-let-7i*	122.8	281.4
hsa-miR-106b	8639.8	18881
hsa-miR-107	725.8	1938.9
hsa-miR-17*	433.5	941.8
hsa-miR-18a	1060.8	2560
hsa-miR-20b	2163.5	5665.8
hsa-miR-214	172.3	383.4
hsa-miR-216a	89.1	197.3
hsa-miR-217	86.3	183.8
hsa-miR-221*	54.3	113.8
hsa-miR-330-3p	213.1	443.2
hsa-miR-452*	189.7	633.3
hsa-miR-509-3-5p	157	371.4
hsa-miR-517*	109.9	230.8
hsa-miR-518e*	88.1	196.4
hsa-miR-519b-5p	72.5	155.1
hsa-miR-593*	175.4	356.7
hsa-miR-621	178.6	486.7
hsa-miR-646	150.9	350.6
hsa-miR-767-5p	107.1	232.4
UP		
hsa-let-7d*	178.6	37.7
hsa-miR-1249	144.8	46.1
hsa-miR-125a-5p	370.8	147.4
hsa-miR-1280	6779.6	2676.2
hsa-miR-142-3p	105.3	2
hsa-miR-145	358	94.6
hsa-miR-146a	326.8	161.8
hsa-miR-151-3p	999	422.6
hsa-miR-181a-2*	154.8	64.7
hsa-miR-183*	195.9	87.7
hsa-miR-186	206.5	26.2
hsa-miR-18a*	397.4	135

Knowledge Representation in Melanoma Cancer

- Using logical representation formalisms
  - **Readability** of the results (IF-THEN rules)
  - **No ambiguity** (YES or NO) physicians may need in melanoma therapy
    - ``relapse(*patient*)'' : the *patient* relapsed.
    - ``died(*patient*)'' : the *patient* died.
    - ``Metastasis(*patient*)'' : Metastasis in the *patient* occured.
- Discretized data according to an expert's opinions
  - 3 statuses: low, medium, high

    ex. ``mir160(*patient,* low)'' : mir160 is lower than in healthy patients

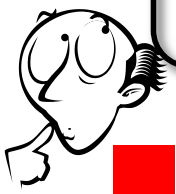
    ``age(*patient*, medium)'' : the age of patient is average

Problem setting

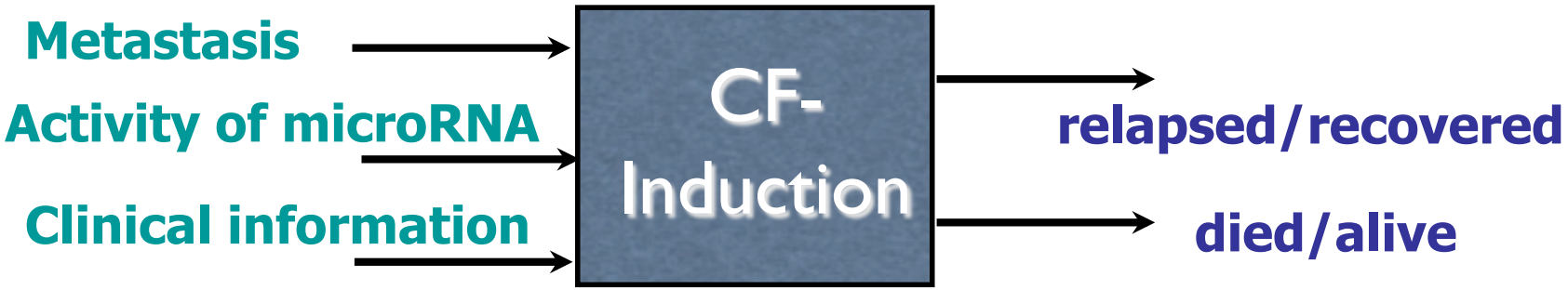
- Observations ***O*** : Information on life-extension
  - **healthy, Melanoma** (*Binary*)
- Background theory ***B*** : Factors related to the emergence of cancer
  - **Metastasis information** (*Binary*): Existence of metastasis
  - **Activity of miRNA** (*Continuous*) :
    - has-mir-7i* 122.8/281.4
    - Discretizing % of expression is measured for each patient and compared with healthy one : **N° Patients M10 and corresponding hsa-let-7b, hsa-miR-17, hsa-miR-18a hsa-miR-20a hsa-miR-21 hsa-miR-34a hsa-miR-130a hsa-miR-141 hsa-miR-143 hsa-miR-145 hsa-miR-146a hsa-miR-152 hsa-miR-155 hsa-miR-185 hsa-miR-191 hsa-miR-200c hsa-miR-221 hsa-miR-222 hsa-miR-338-3p hsa-miR-1246 hsa-miR-1290 hsa-miR-2110 miR-27b**
  - **Daily/General clinical information** (*Continuous*):
    - Age, Breslow index, Ulceration

- Goal:

Extracting **logical causal relations**
between those **19 factors** and **life-extension**



- Approach: Using CF-Induction



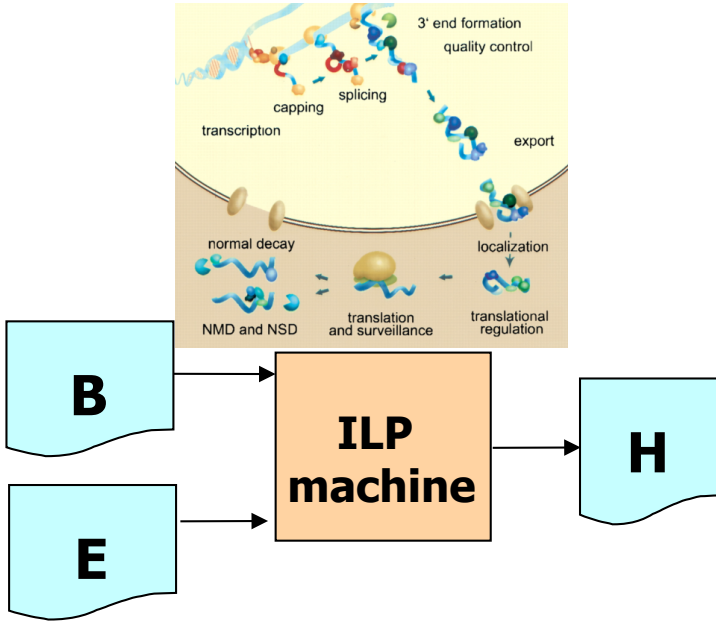
Knowledge Based Discovery

A **logic-based machine learning** technique

— Richer representation formalisms (First-order predicate logic)

— Classification

- **Input:**
  - ***B*** : background theory.
  - ***E*** : observations.
  - ***LB*** : language bias for restricting the syntax of hypotheses
- **Output:**
  - *H* : hypothesis satisfying that
    - *H is a clause belonging to LB*
    - *B* ∧ *H* logically explains *E*



Background

Breslow(X,high)∧mir21(X,high) ∧miR-222, miR-23a(X,high) ∧miR-92(X,high) ∧ miR-149(X,high) ∧ miR-221 (X,high)➔ relapse(X)

CF-Induction [Inoue, 2001; 2004]

$$B \wedge H \models E$$

$$\Leftrightarrow \quad B \wedge \neg E \models \neg H$$

- Based on **Inverse Entailment** like Progol
- Compute the ***characteristic clauses*** of $B \wedge \neg E$ using a ***consequence-finding*** procedure (SOLAR).
- Includes the bottom method and abductive computation.
  - B : full clausal theory (non-Horn clauses)
  - E : full clausal theory (non-Horn clauses)
  - H : full clausal theory (non-Horn clauses)
- Sound and complete

CF-Induction: Principle

$B \wedge H \models E$

$\Leftrightarrow B \wedge \neg E \models \neg H$

$\Leftrightarrow B \wedge \neg E \models Carc(B \wedge \neg E, \mathbf{P}) \models CC(B,E) \models \neg H$

$\Leftrightarrow CC(B,E) \subseteq Carc(B \wedge \neg E, \mathbf{P})$,

$\neg CC(B,E) \equiv F, \quad H \models F$ (where F is CNF)

CF-Induction: Algorithm

1. Compute $Carc(B \wedge \neg E, \boldsymbol{P})$.
2. Construct $CC(B,E)$ such that
   - $CC(B,E) \subseteq Carc(B \wedge \neg E, \boldsymbol{P})$;
   - $CC(B,E) \cap NewCarc(B, \neg E, \boldsymbol{P}) \neq \varphi$.
3. Convert $\neg CC(B,E)$ into CNF F .
4. Generalize F to H such that
   - $B \wedge H$ is consistent.

Results with our ILP system

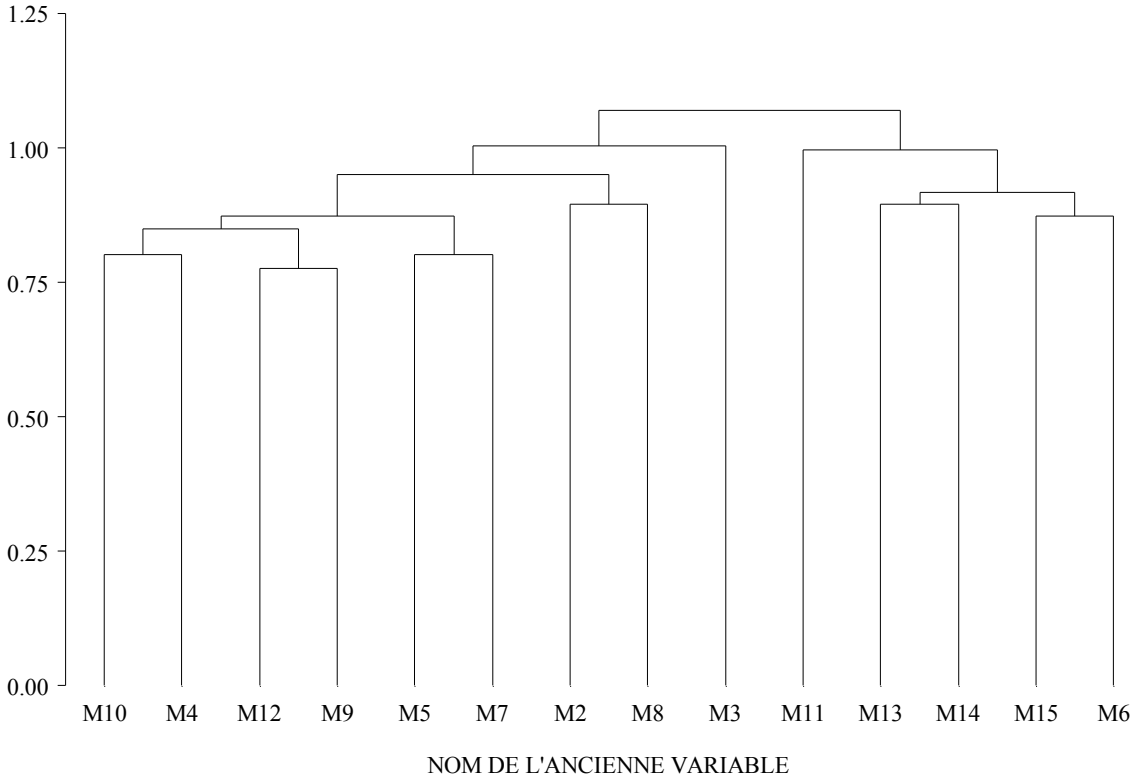
- Input file: the information on **15 patients with melanoma / 5 healthy patients**
- Causal relations between status of **relapsing** and **19 factors**

Hypothesis (Clause)	**Rs** (%)	**Rf** (%)	**Expert's opinion**
(7) [mir182(X, low), metastasis(X), age(X, medium) → relapse(X)]	50	8	○
(6) [metastasis(X), age(X, medium) → relapse(X)]	40	10	○
(7) **[mir630(X, low), mir182(X,high) → relapse(X)]**	**30**	**10**	

Conclusion

We identified a group of patients(M6, M15, M14) with a rapid evolution of melanoma.

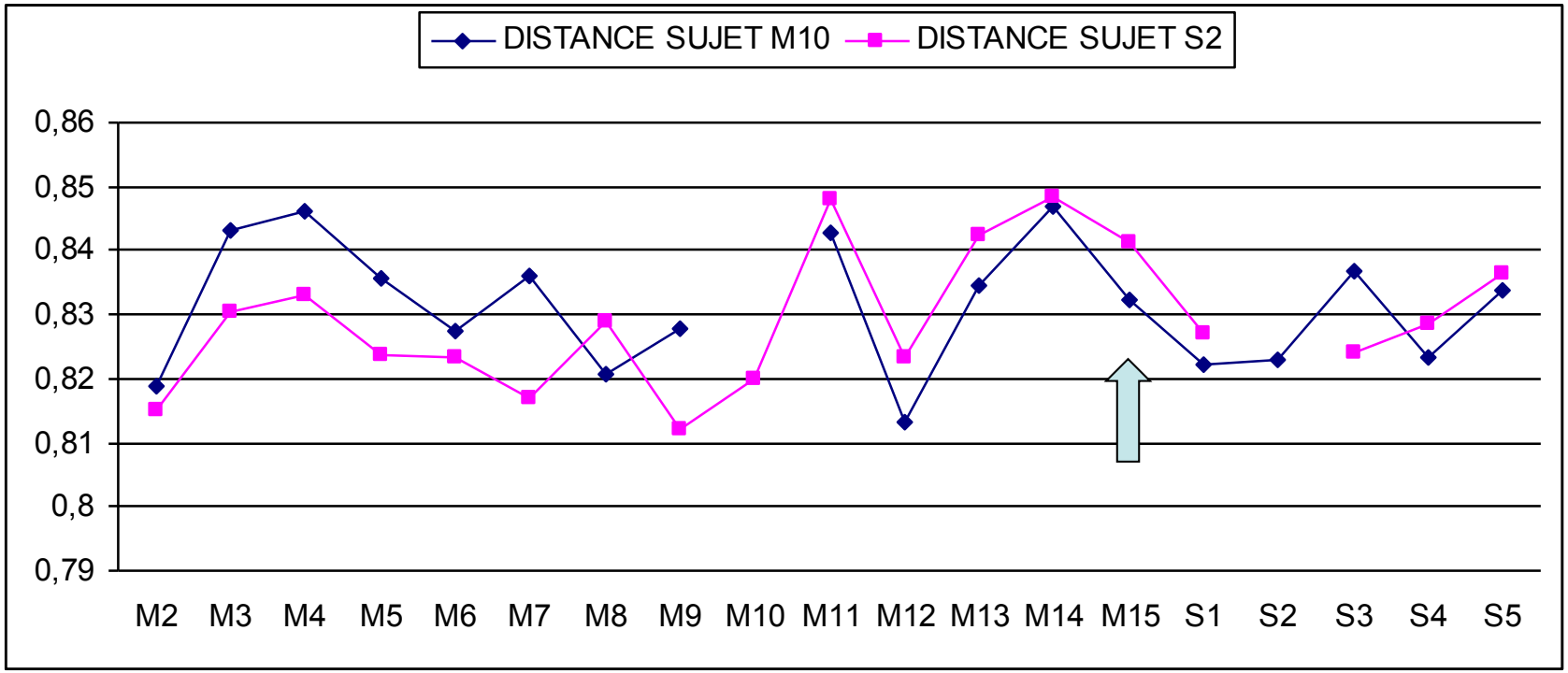
[mir630(X, low), mir182(X,high)]



53 microRNA

Conclusion

Kolmogorov Complexity shows a relatively important distance
Between patient M15 and a healthy one (S2)



Thank you

MicroRNA Expression Patterns to Differentiate Pancreatic Adenocarcinoma From Normal Pancreas and Chronic Pancreatitis.

Bloomston, Mark; Frankel, Wendy; Petrocca, Fabio; Volinia, Stefano; Alder, Hansjuerg; Hagan, John; Liu, Chang-Gong; Bhatt, Darshna; Taccioli, Cristian; Croce, Carlo

JAMA. 297, 1901-1908 (2007)

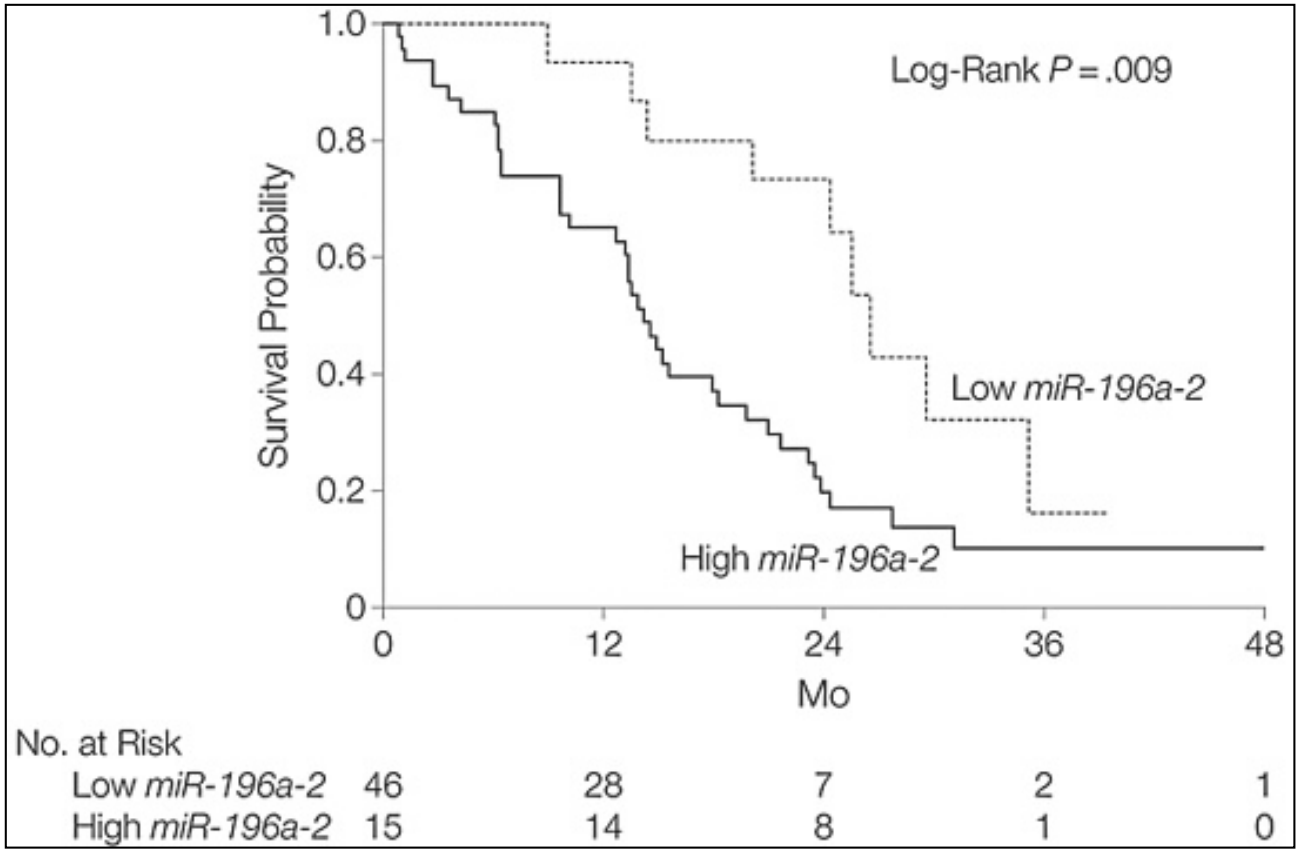


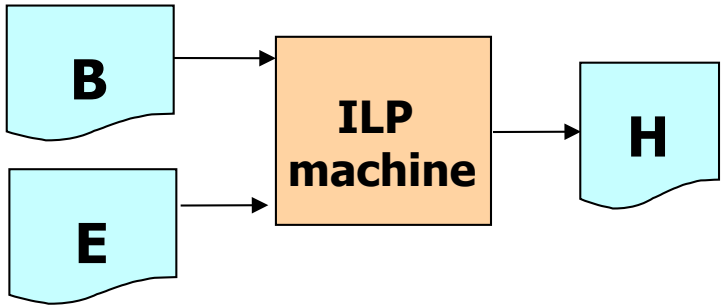
Figure 4 . Kaplan-Meier Overall Survival Curve for Patients With Pancreatic Cancer, Based on Expression of miR-196-a2

Inductive Logic Programming:

A **logic-based machine learning** technique

- — Richer representation formalisms (First-order predicate logic)
- — Classification

- **Input:**
  - ***B*** : background theory.
  - ***E*** : observations.
  - ***LB*** : language bias for restricting the syntax of hypotheses



- **Output:**
  - *H* : hypothesis satisfying that
    - *H is a clause belonging to LB*
    - *B* ∧ *H* logically explains *E*

Comparison with other methods

Results with our ILP system 1/2

- Input file: the information on **15 patients with melanoma / 5 healthy patients**
- Causal relations between status of **relapsing** and **10 facto**

Hypothesis (Clause) Ts: 50 (%), Tf: 50 (%) (Ranking using Rs)	**Rs** (%)	**Rf** (%)	**Expert's opinion**
(7) [hnRNPA1(X, high), n(X), age(X, medium) → relapse(X)]	57	8	○
(6) [n(X), age(X, medium) → relapse(X)]	62	10	○
(7) [gb9(X, high), hnRNPA1(high), n(X) → relapse(X)]	**57**	**10**	
(4) [hnRNPA1(X, high), n(X) → relapse(X)]	69	10	○
(7) [gb9(X, high), n(X) → relapse(X)]	**57**	**11**	
(2) [n(X) → relapse(X)]	74	11	○
(10) [gb9(X, high), hnRNPA1(X, high), age(X, medium) → relapse(X)]	**55**	**22**	
(10) [gb9(X, high), age(X, medium) → relapse(X)]	**55**	**25**	
(2) [hnRNPA1(X, high), age(X, medium) → relapse(X)]	74	30	○
(5) [gb9(X, high) → relapse(X)]	**67**	**35**	
(1) [age(X, medium) → relapse(X)]	81	38	○
(14) [hnRNPA1(X, high), aSF_SF2(X, high) → relapse(X)]	50	38	○
(10) [pr(X) → relapse(X)]	55	45	○

System description

- **Target hypotheses**: Let ***LB*** be a language bias, ***B*** a background theory, ***O*** observations, ***Tp*** a positive threshold and ***Tn*** be a negative threshold. If a clause *H* satisfies the following conditions, then *H* be a ***hypothesis*** with respect to *LB*, *B*, *O*, *Tp* and *Tn*:
  1. *H* belongs to *LB*;
  2. $|Rs| \geqq Ts$, where $|Rs|$ is ratio (%) of observations that can be explained by $B \wedge H$;
  3. $|Rf| \leqq Tf$, where $|Rf|$ is ratio (%) of observations that are inconsistent with $B \wedge H$.
- Our ILP system can enumerate **all the target hypotheses**.
- Note that the search strategy is based on top-down approach.

Results with our ILP system 2/2

- Causal relations between status of **recovering** and **10 factors**

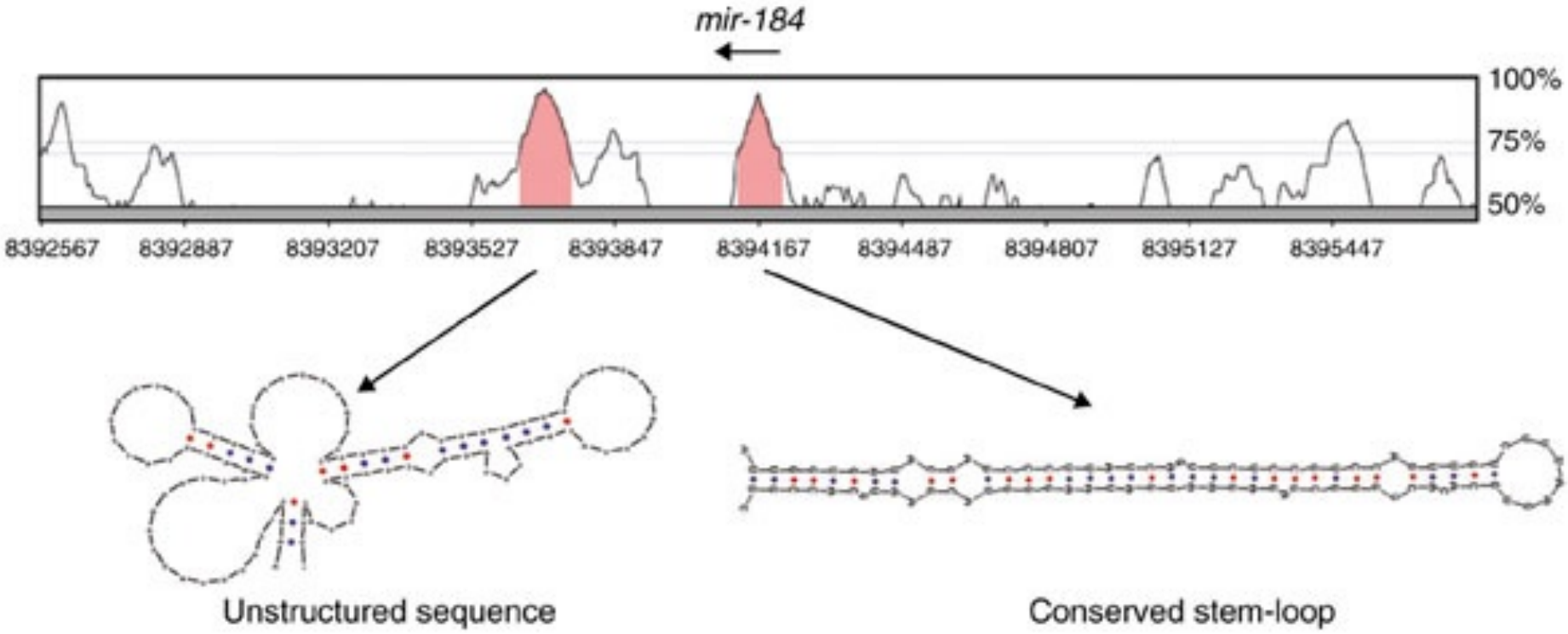
Hypothesis (Clause) Ts: 50 (%), Tf: 50 (%) (Ranking using Rs)	**Rs** (%)	**Rf** (%)	**Expert's opinion**
(5) [aSF_SF2(X, high), not n(X) → recover(X)]	**55**	**22**	
(6) [hnRNPA1(X, high), er(X), not n(X) → recover(X)]	**50**	**22**	
(7) [er(X), pr(X,), not n(X), chemotherapy(X → recover(X)]	50	25	○
(8) [pr(X), not n(X), chemotherapy(X) → recover(X)]	**50**	**25**	
(9) [er(X), not n(X), chemotherapy(X) → recover(X)]	**50**	**25**	
(10) [not n(X), chemotherapy(X) → recover(X)]	**50**	**25**	
(11) [er(X), pr(X), chemotherapy(X) → recover(X)]	50	25	○
(12) [pr(X), chemotherapy(X) → recover(X)]	50	25	○
(13) [er(X), chemotherapy(X) → recover(X)]	50	25	○
(14) [chemotherapy(X) → recover(X)]	**50**	**25**	
(15) [er(X, on), pr(X), not n(X) → recover(X)]	50	25	○
(3) [er(X), not n(X) → recover(X)]	**60**	**30**	
(2) [hnRNPA1(X, high), not n(X) → recver(X)]	**65**	**32**	
(4) [pr(X), not n(X) → recover(X)]	**57**	**32**	
(1) [not n(X) → recover(X)]	85	40	○

Comparative genomics

- Start with 24 known Drosophila pre-miRNAs (the ~70-100 long transcripts before miRNAs)

- All are found to be conserved beween D. melanogaster and D. pseudoobscura
  - Typically, more conserved than gene. (The third codon “wobble” not relevant here)

http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=12844358



miRNA genes are isolated, evolutionarily conserved genomic sequences that have the capacity to form extended stem-loop structures as RNA. Shown are VISTA plots of globally aligned sequence from D. melanogaster and D. pseudoobscura, in which the degree of conservation is represented by the height of the peak. This particular region contains a conserved sequence identified in this study that adopts a stem-loop structure characteristic of known miRNAs. Expression of this sequence was confirmed by northern analysis (Table 2), and it was subsequently determined to be the fly ortholog of mammalian mir-184. Most conserved sequences do not have the ability to form extended stem-loops, as evidenced by the fold adopted by the sequence in the neighboring peak.

Finding microRNA genes

- Find highly conserved sequences, length ~70-100
- Check for secondary structure
- Are we done?
  - No, too many such sequences; more filters needed

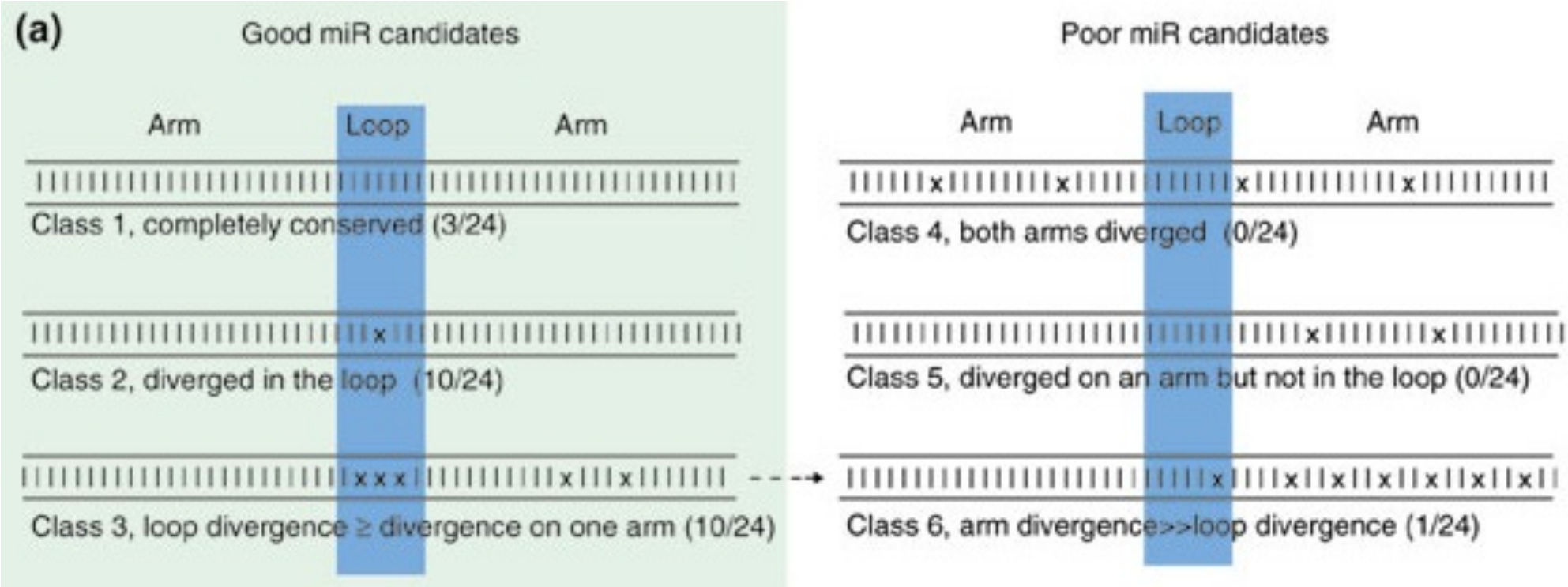
Comparative genomics

- Look carefully at pairwise alignments of each of the 24 pairs or orthologous pre-miRNAs.
- Only three pairs completely conserved
- Ten pairs are diverged exclusively within their loop sequence;no pair diverged exclusively in arm
- Of the 11 remaining, seven show more changes in the loop than in non-miRNA-encoding arm

How to find miRNAs?

- Experimental methods so far
- Lai et al (2003) one of the works that try solving this problem computationally
- Basic idea:
  - look for evolutionarily conserved sequences
  - check if some of these fold well into the stem-loop structure (“hairpins”) associated with miRNAs

http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=12844358



So what do we learn?

- That class 1 - 3 are the normal pattern of evolutionary divergence of miRNAs
- That classes 4 - 6 are unlikely
- Therefore use these criteria as additional filters for evolutionarily conserved sequences

Prediction Pipeline details: 1

- Align the two genomes
- “Regions” that should contain miRNA genes are estimated as those having
  - length 100,
  - <= 15% mismatches,
  - <= 13% gaps

Pipeline details: 2

- Analyze conserved regions with mfold3.1, an RNA folding algorithm

- Find the top scoring regions (from the mfold program) -- these are candidates for the next stage

Pipeline details: 3

- Assess the divergence pattern of candidate miRNAs

- Boolean filters: remove candidates with
  - exclusive divergence in arm
  - more divergence in miRNA-coding arm than in loop

Final results

- 200 candidate miRNAs came out
- Experimental validation of many of these
- 24 novel miRNAs confirmed

Summary of part 1

- Learned what miRNAs are
- and how the genes encoding these are predicted computationally
- Learned that the miRNAs function to regulated gene expression by binding to the mRNA of the target genes (perfectly or imperfectly)

Part 2: finding the targets

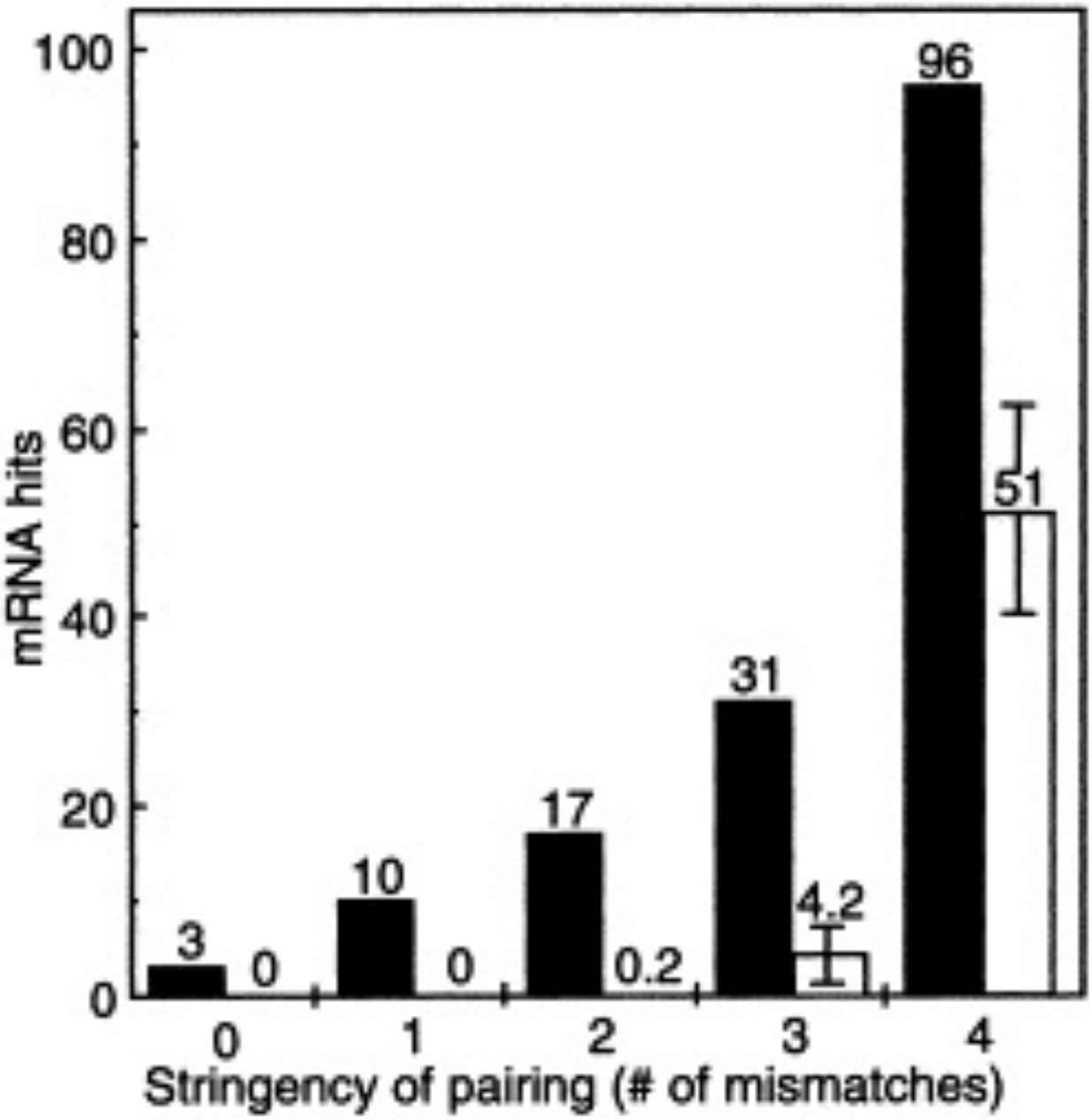
- Rhoades et al (2002)
- We should be looking for targets …
- … with base complementarity
- But small size (20-24 nt) and imperfect base pairing imply that we may ending up predicting too many
- Rhoades et al found that nearly perfect complementarity is a good indicator of miRNA targets in plant

Plant miRNAs

- Started with 16 known *Arabidopsis* miRNAs
- Looked for complementary strings with <= 4 mismatches and no gaps
- Also did the same genome-wide search with “randomized” versions of the 16 miRNAs

doi:10.1016/S0092-8674(02)00863-2

Results of this scan



Near perfect complementarity

- Number of hits with <= 3 mismatches is 30 for the real miRNAs, 0.2 for the random
  - Why fractional for random?
- Therefore <= 3 matches supposed to be a good indicator of targets
- Find all targets using this rule; as simple as that!

Alternative Splicing

(a review by Liliana Florea, 2006)

What is alternative splicing?

- The first result of transcription is “pre-mRNA”
- This undergoes “splicing”, i.e., introns are excised out, and exons remain, to form mRNA
- This splicing process may involve different combinations of exons, leading to different mRNAs, and different proteins
- This is alternative splicing

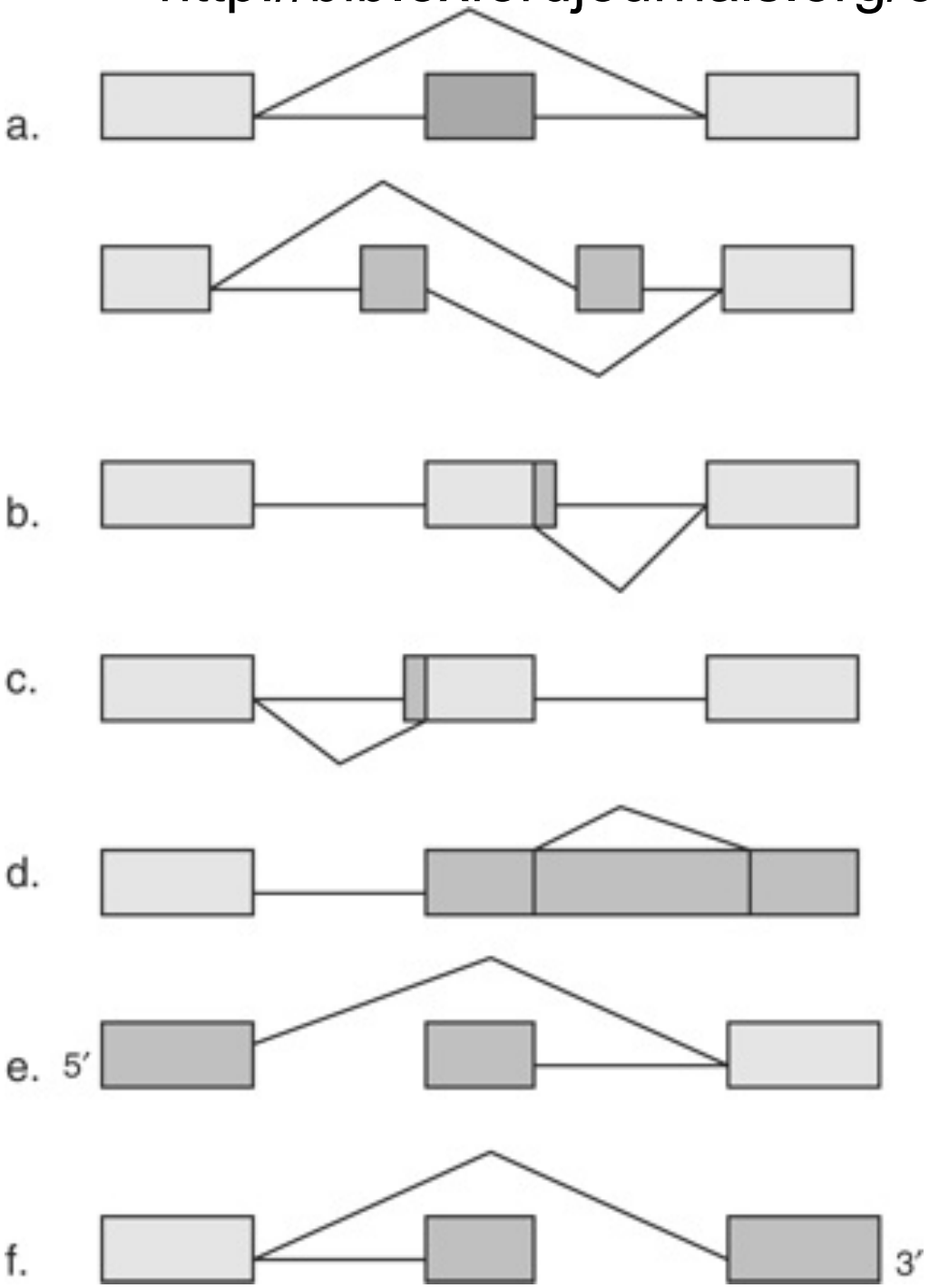
Significance

- Important regulatory mechanism, for modulating gene and protein content in the cell
- Large-scale genomic data today suggests that as many as 60% of the human genes undergo alternative splicing

Significance

- Number of human genes has recently been estimated to be about 20-25 K.
- Not significantly greater than much less complex organisms
- Alternative splicing is a potential explanation of how a large variety of proteins can be achieved with a small number of genes
- Errors in splicing mechanism implicated in diseases such as cancers

http://bib.oxfordjournals.org/cgi/content/full/7/1/55/F1



exon inclusion/exclusion

alternative 3’ exon end

alternative 5’ exon end

intron retention

5’ alternative UTR

3’ alternative UTR

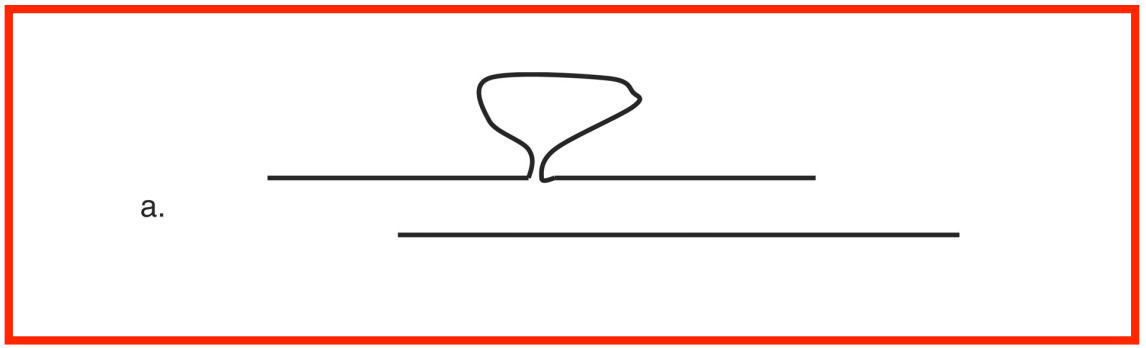
Bioinformatics of Alt. splicing

- Two main goals:
  - Find out cases of alt. splicing
    - What are the different forms (“isoforms”) of a gene?
  - Find out how alt. splicing is regulated
    - What are the sequence motifs controlling alt. splicing, and deciding which isoform will be produced

Identification of splice variants

- Direct comparison between sequences of different cDNA isoforms
  - Q: What is cDNA? How is this different from a gene's DNA?
  - cDNA is "complementary DNA", obtained by reverse transcription from mRNA. It has no introns
- Direct comparison reveals differences in the isoforms
- But this difference could be part of an exon, a whole exon, or a set of exons

Bioinformatics methods for identifying alternative splicing



direct comparison

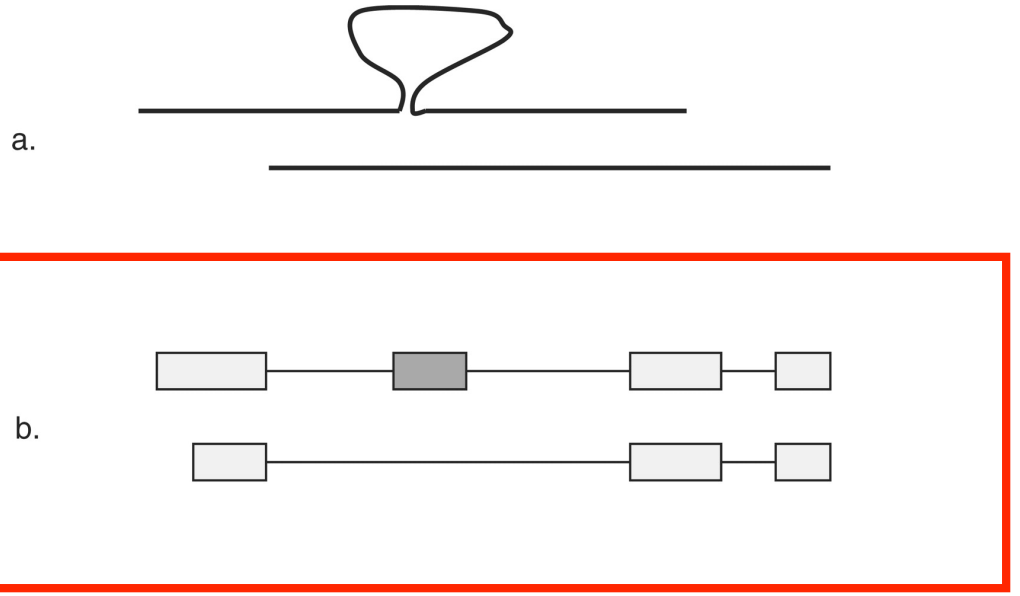
Florea, L. Brief Bioinform 2006 7:55-69; doi:10.1093/bib/bbk005

Briefings in Bioinformatics

Identification of splice variants

- Comparison of exon-intron structures (the gene's architecture)
- Where do the exon-intron structures come from?
  - Align cDNA (no introns) with genomic sequence (with introns)
  - This gives us the intron and exon structure

Bioinformatics methods for identifying alternative splicing



comparison of exon-intron structures

Florea, L. Brief Bioinform 2006 7:55-69; doi:10.1093/bib/bbk005

Briefings in Bioinformatics

Identification of splice variants

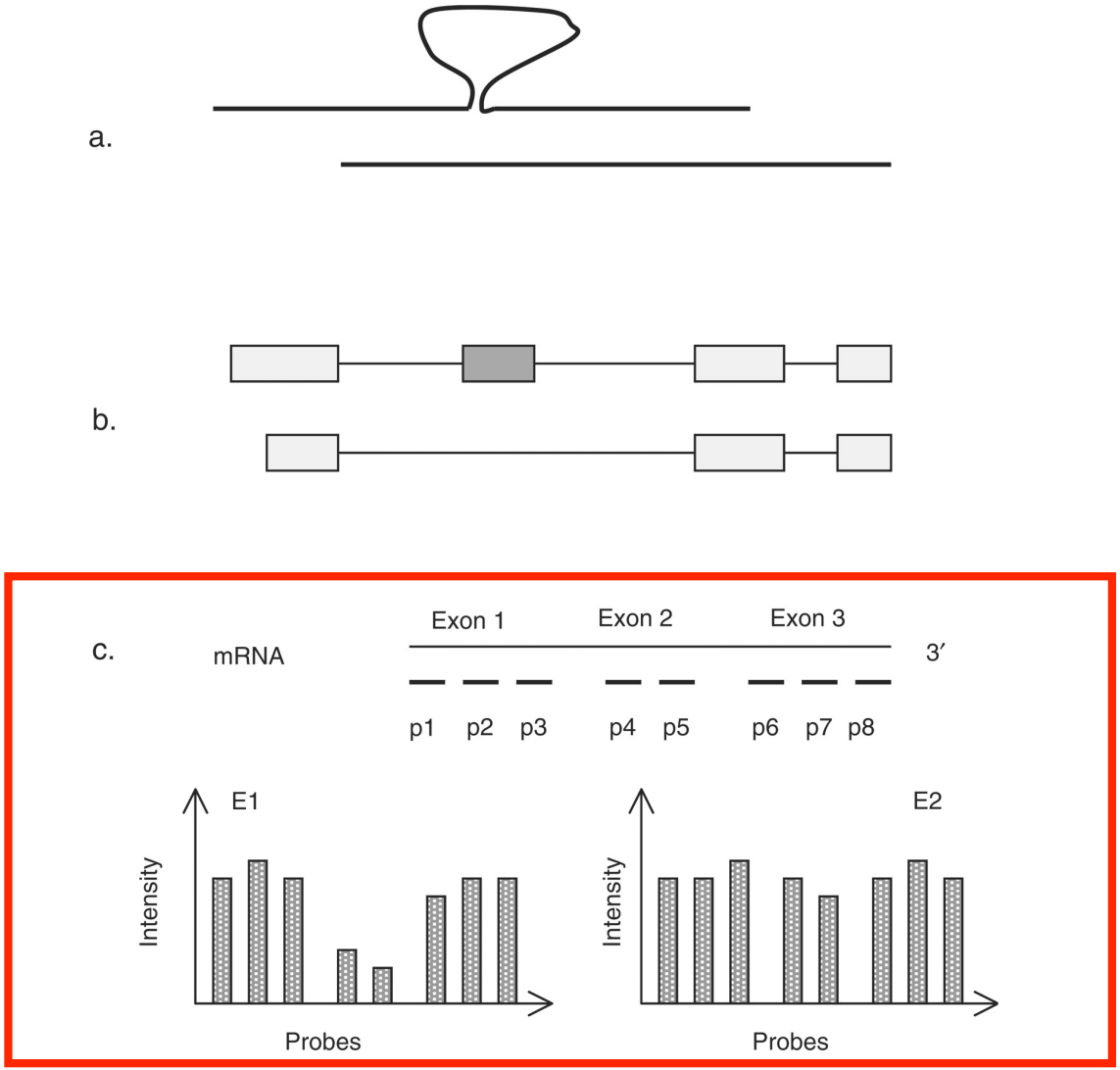
- Alignment tools.
- Align cDNA sequence to genomic sequence
- Why shouldn't this be a perfect match with gaps (introns)?
  - Sequencing errors, polymorphisms, etc.
- Special purpose alignment programs for this purpose

Splice variants from microarray data

- Affymetrix GeneChip technology uses 22 probes collected from exons or straddling exon boundaries
- When an exon is alternatively spliced, expression level of its probes will be different in different experiments

Bioinformatics methods for identifying alternative splicing

splice variants from micro array data



Florea, L. Brief Bioinform 2006 7:55-69; doi:10.1093/bib/bbk005

Briefings in Bioinformatics

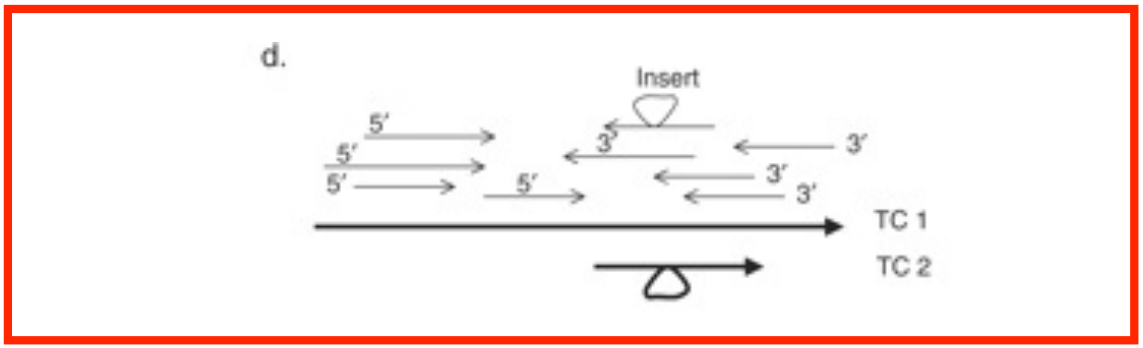
Identifying full lengh alt. spliced transcripts

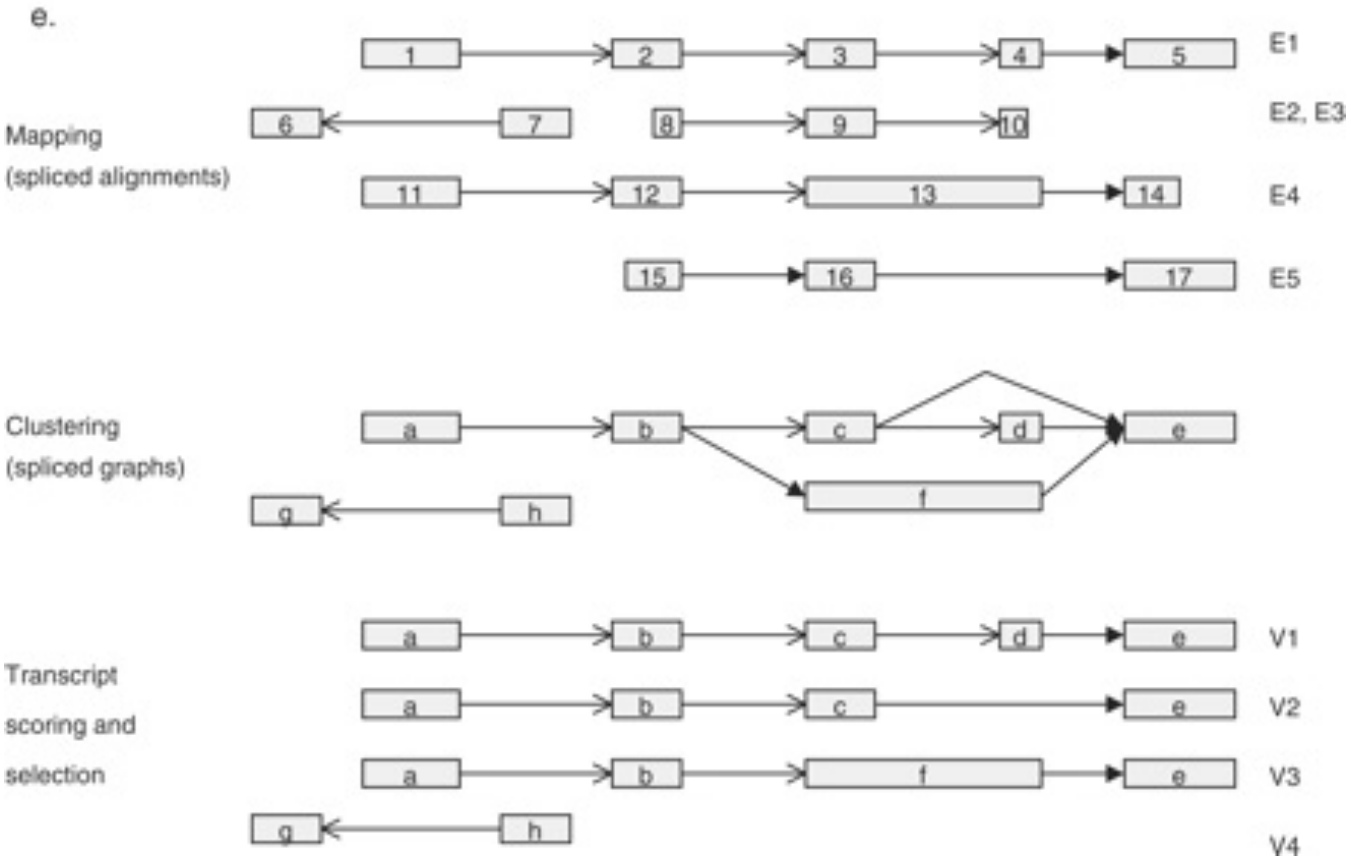
- Previous methods identified parts of alt. spliced transcript
- We assumed we had access to the cDNA sequence, i.e., the full transcript
- Much more difficult to identify full length transcripts (i.e., all alt. spliced forms)

Method 1 (“gene indices”)

- EST is the sequence of a partial transcript
- Compare all EST sequences against one another
- Identify significant overlaps
- Group and assemble sequences with compatible overlaps into clusters
- Similar to the assembly task, except that we are also dealing with alt. spliced forms here

Gene indices





Problems with this method

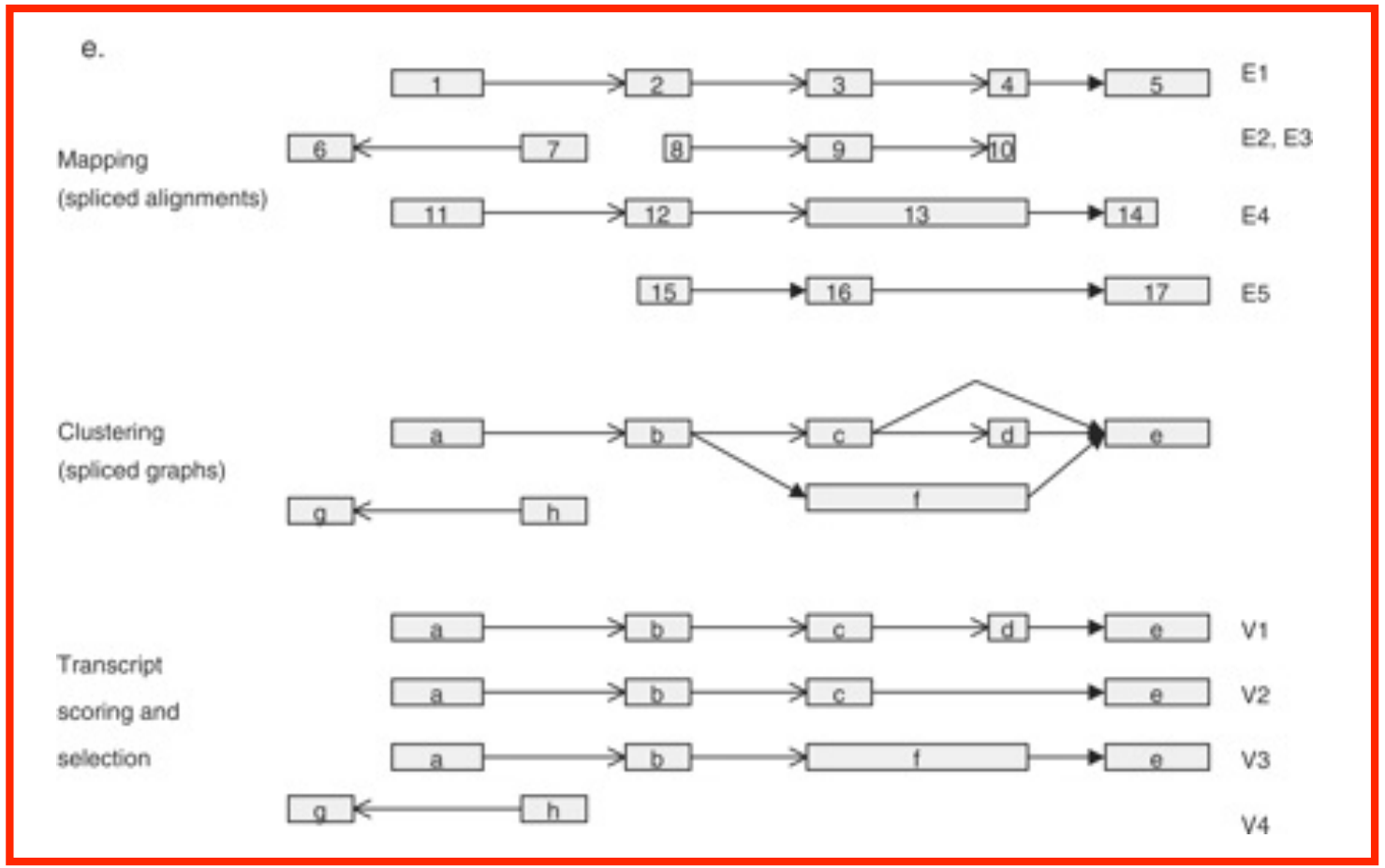
- Overclustering: paralogs may get clustered together.
  - What are paralogs?
  - Related but distinct genes in the same species
- Underclustering: if number of ESTs is not sufficient
- Computationally expensive:
  - Quadratic time complexity

Method 2: Splice graphs

- Nodes: Exons
- Edges: Introns
- Gene: directed acyclic graph
- Each path in this DAG is an alternative transcript

Spliced alignments of cDNAs on the genome (E1–E5) are clustered along the genomic axis and consolidated into splice graphs. Vertices in the splice graph represent exons (a–h), arcs are introns connecting the exons consistently with the cDNA evidence, and a branching in the graph signals an alternative splicing event. Splice variants (V1–V4) are read from the graph as paths from a source vertex (with no 'in' arc) to a sink vertex (with no 'out' arc).

Splice graph



Splice graphs

- Combinatorially generate all possible alt. transcripts
- But not all such transcripts are going to be present
- Need scores for candidate transcripts, in order to differentiate between the biologically relevant ones and the artifactual ones

Summary

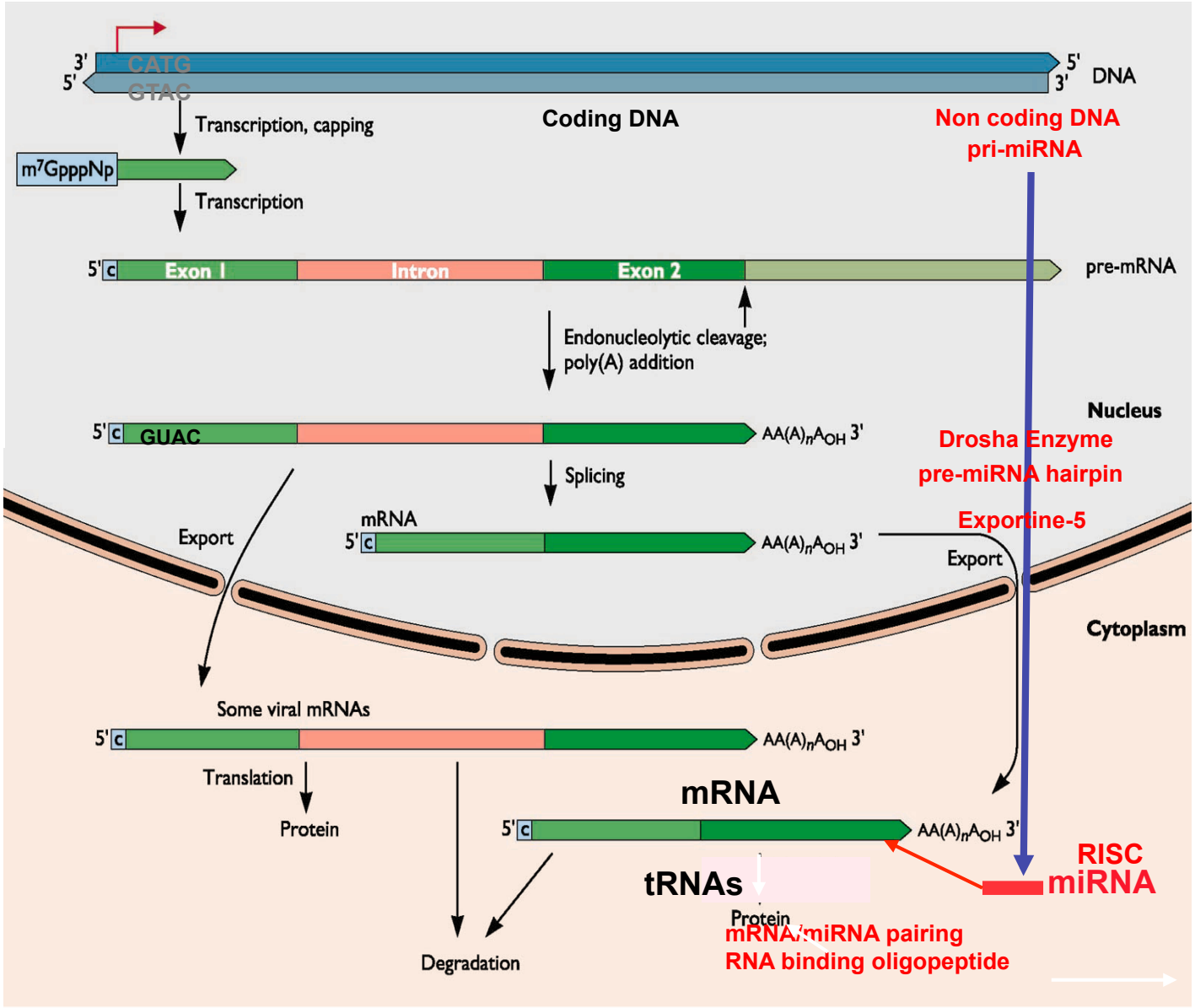
- Alternative splicing is very important
- Bioinformatics for finding alternative spliced forms

Gene Expression

- Process in which a gene convert the coded information stored in its DNA sequence into essential proteins, which are needed to perform and regulate most basic function.

- Expression is often related to mRNA through which the protein-coding instruction from the gene are transmitted.

Overview of mRNA and miRNA Processing



Microarray chips

- …microarrays can measure many genes at once.
- Microarray chips are commonly glass slides with a matrix of spots printed (using eg. dot matrix technology) on to them.
- A spot contains millions of identical molecules of DNA or **oligonucleotide** (the **probes**), which will bind a specific DNA sequence, such as the cDNA of a gene.
- The glass slides can contain 1000s of spots, each recognising a different sequence, eg. one spot for every gene in the human genome.

Microarray experiments

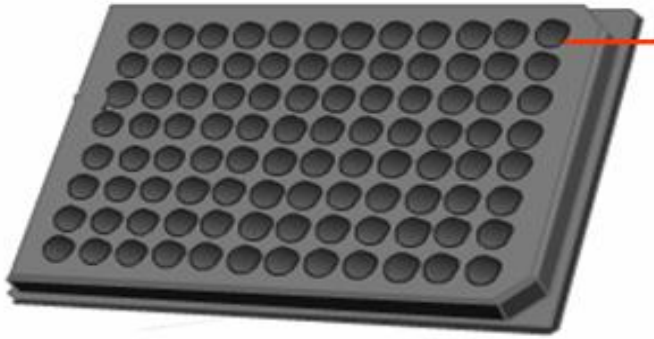
- Since almost all mRNA $\xrightarrow{\text{translated}}$ protein, total mRNA of cell ~ genes expressed.
- Mash up cells and extract mRNA.
- Reverse transcribe RNA → cDNA (can be heated to make single-stranded).
- Label cDNA from **reference** cells green (Cy3) and cDNA from **target** cells red (Cy5).
- Hybridise (wash on equal amounts of target & reference sample & allow to bind to probes which have complementary bases) both samples, reference and target, to a single microarray chip.

Results of microarray experiments

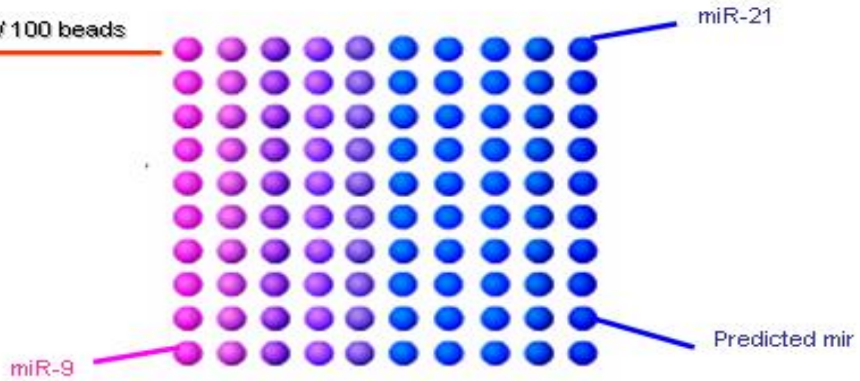
- The spot for gene 1 =
  - red if more mRNA 1 in target cells
  - green if more mRNA 1 in reference cells
  - yellow if same in both
- Actually, images of red & green fluorescence are taken separately using laser & scanner & their intensities are measured using image software.
- Data often expressed as matrix of relative expression levels = $\frac{\text{intensity red}}{\text{intensity green}}$ indexed by genes and target samples.

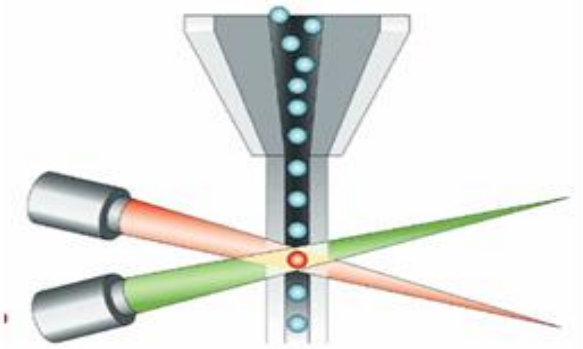
The LUMINEX Technology

1 sample / 100 microRNA tested



1 well / 100 beads

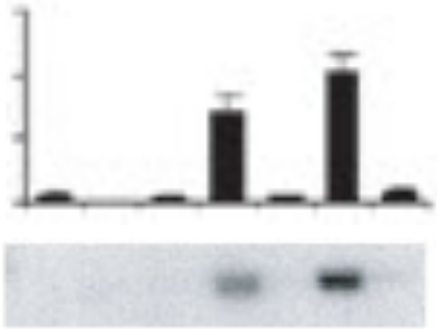


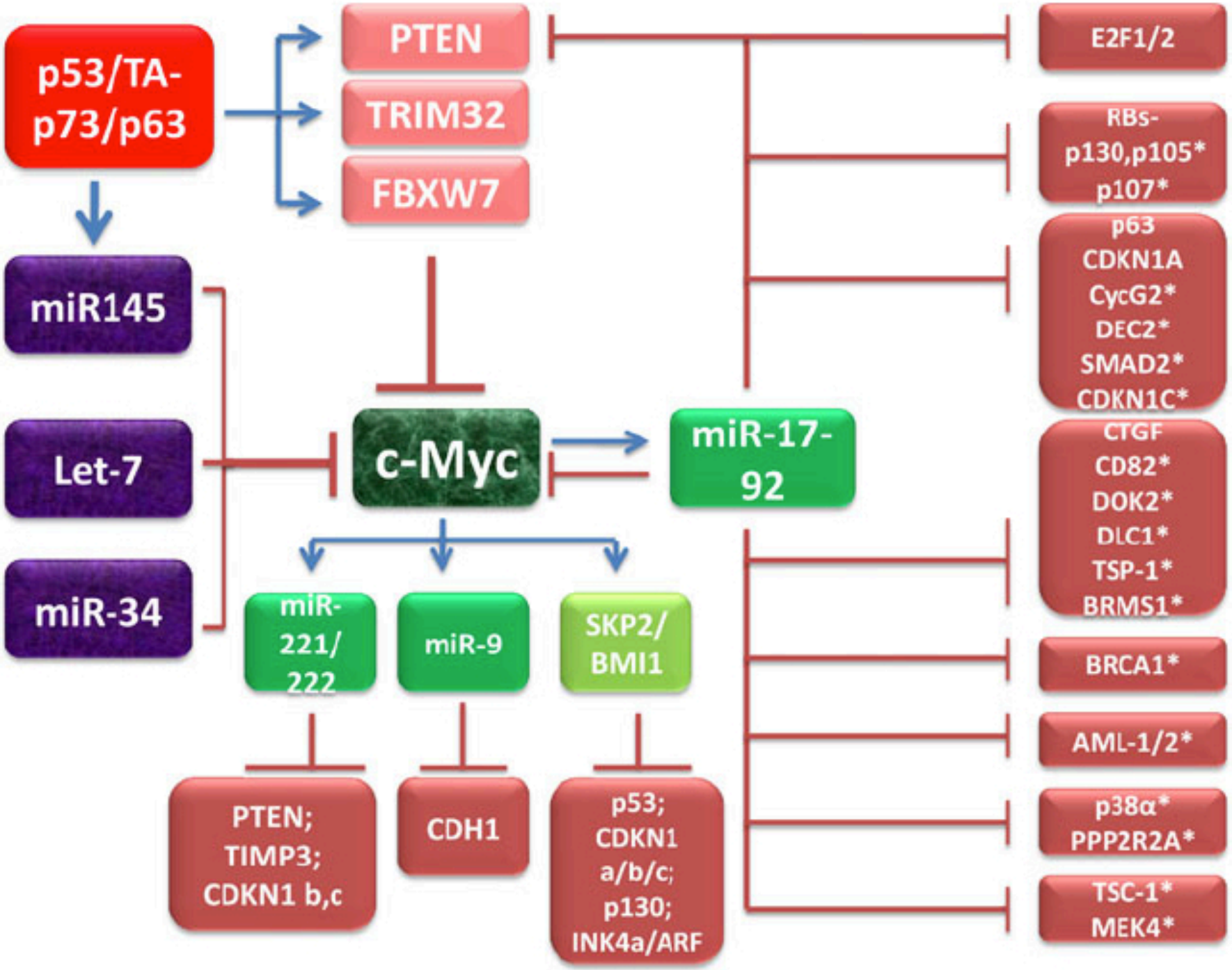


Two laser to detect the bead and the labeled microRNAs

1 bead / 1 microRNA

mir221





L. Boominathan Cancer Metastasis Rev 29 (2010)