Algorithmic Challenges from New Sequencing Technologies

Sven Rahmann

Bioinformatics for High-Throughput Technologies
Algorithm Engineering, Computer Science 11
TU Dortmund

JOBIM 2010, Montpellier

Part 1

Sequencing Technologies and Read Mapping

“New” Sequencing Technologies...

...produce a flood of short DNA reads from pepared sample(s).

- Roche 454
- Illumina Solexa
- ABI SOLiD
- Ion Torrent
- Pacific Biosciences
- upcoming single molecule technologies ...

(This talk: no details on these technologies)

Example

- Illumina HiSeq 2000: 20–25 Gbp / day
- Beijing Genomics Institute ordered 128 of them
- Output: $\approx$ 3 Tbp / day (1000x Human Genome)

Applications & Replacement of Other Technologies

Genomics

- De novo sequencing
- Resequencing: variation discovery
  (e.g., SNP and CNV discovery)
- Description of the pan-genome of a species

Applications & Replacement of Other Technologies

Genomics

- De novo sequencing
- Resequencing: variation discovery
  (e.g., SNP and CNV discovery)
- Description of the pan-genome of a species

Transcriptomics

- Discovery of full transcriptome
- Gene / exon / (small) RNA expression analysis:
  mRNA-Seq, DGE, SuperSAGE

Applications & Replacement of Other Technologies

Epigenomics

- Determination of DNA methylation state

Applications & Replacement of Other Technologies

Epigenomics

- Determination of DNA methylation state

Metagenomics

- Genomic composition of ecological communities
  $\Rightarrow$ Implications for inter-species relations

Applications & Replacement of Other Technologies

Epigenomics

- Determination of DNA methylation state

Metagenomics

- Genomic composition of ecological communities
  $\Rightarrow$ Implications for inter-species relations

Transcriptional regulation

- ChIP-seq $\Rightarrow$ transcription factor binding motifs

Challenge

Understand and Use Internal Sequencer State

- Presently, data consists of:
  1. DNA reads (base calls)
  2. Quality values

Challenge

Understand and Use Internal Sequencer State

- Presently, data consists of:
  1. DNA reads (base calls)
  2. Quality values
- However, sequencer has more information.
- Example: SOLiD Color Space encodes dinucleotides
  TTACGG is T,TT,TA,AC,CG,GG = TXXX

Challenge

Understand and Use Internal Sequencer State

- Presently, data consists of:
  1. DNA reads (base calls)
  2. Quality values
- However, sequencer has more information.
- Example: SOLiD Color Space encodes dinucleotides
  TTACGG is T,TT,TA,AC,CG,GG = TXXX
- Needed: Standards to encode machine state information
- Analysis based on this information (more than DNA+quality)

(D. Haussler, HiTSeq 2010

Read Mapping: A Fundamental Task

Read Mapping Problem

Given:

- short DNA sequence read (string),
- per-base quality values
- reference sequence (string)
- error rate threshold

Read Mapping: A Fundamental Task

Read Mapping Problem

Given:
- short DNA sequence read (string),
- per-base quality values
- reference sequence (string)
- error rate threshold

Sought:
- all/one location(s) in reference where read occurs with (quality-weighted) error rate below threshold ($\approx$ classical approximate matching / alignment.)

Read Mapping: A Fundamental Task

Read Mapping Problem

Given:
- short DNA sequence read (string),
- per-base quality values
- reference sequence (string)
- error rate threshold

Sought:
- all/one location(s) in reference where read occurs with (quality-weighted) error rate below threshold ($\approx$ classical approximate matching / alignment.)

Variations:
- local instead of semiglobal,
- read ends may be adapters,
- spliced alignment across exon boundaries

Challenges & Issues with State of the Art

- Speed of read mapping: index necessary
  1. index (parts of) the reference
  2. index (parts of) the reads

Challenges & Issues with State of the Art

- Speed of read mapping: index necessary
  1. index (parts of) the reference
  2. index (parts of) the reads
- Size of reference & data (and indexes): GBs!
  32 bits limit us to addressing 2 GB,
  suffix array of human genome needs 24 GB (uncompressed).

Challenges & Issues with State of the Art

- Speed of read mapping: index necessary
  1. index (parts of) the reference
  2. index (parts of) the reads
- Size of reference & data (and indexes): GBs!
  32 bits limit us to addressing 2 GB,
  suffix array of human genome needs 24 GB (uncompressed).
- Guarantees of read mapping (exact vs. heuristic)

Challenges & Issues with State of the Art

- Speed of read mapping: index necessary
  1. index (parts of) the reference
  2. index (parts of) the reads
- Size of reference & data (and indexes): GBs!
  32 bits limit us to addressing 2 GB,
  suffix array of human genome needs 24 GB (uncompressed).
- Guarantees of read mapping (exact vs. heuristic)
- Dealing with multiple matching loci
  (one best, all best, all suboptimal)
  effects on downstream analysis; repeats.

Indexing

Indexing the Reference

Advantages:

- Reference does not change over time.
- Reference is now smaller than reads.
- Reads must be considered sequentially anyway.

Indexing

Indexing the Reference

Advantages:

- Reference does not change over time.
- Reference is now smaller than reads.
- Reads must be considered sequentially anyway.

Frequently used index structures:

- q-gram index
- (enhanced, extended) suffix array
- compressed self-index

Indexing

Indexing the Reference

Advantages:

- Reference does not change over time.
- Reference is now smaller than reads.
- Reads must be considered sequentially anyway.

Frequently used index structures:

- q-gram index
- (enhanced, extended) suffix array
- compressed self-index

Indexing the Reads

Scan over reference (genome).
Frequently used index structures:

- Automaton to recognize substrings of several reads

Indexing for Short Exact Matches

Indexing helps to locate **exact matches** of read substrings.
Filtration idea: Appropriate choice of q implies:

- No exact q-gram match $\Rightarrow$ no good alignment
- However: exact q-gram match $\not\Rightarrow$ good alignment
- Necessary to verify matches with full alignment.

q-gram index

- Read a q-gram as a base-4 number with q-digits
- Example: AGTTCA $\mapsto$
  $(023310)_4 = 0 \cdot 1 + 1 \cdot 4 + 3 \cdot 16 + 3 \cdot 64 + 2 \cdot 256 + 0 \cdot 1024 = 756$
- 1-to-1 correspondence: q-grams $\leftrightarrow$ integers $\{0, \ldots, 4^q - 1\}$

Alternative: Hashing of (longer) substrings (not 1-to-1)

Challenge

1-Mismatch and 1-Difference Mapping

Develop an index that, for a given string (q-gram), locates all matches with

- Hamming distance ≤ 1 or edit distance ≤ 1
- rapidly
- without wasting memory
- with few false hits when hashing

Challenge

1-Mismatch and 1-Difference Mapping

Develop an index that, for a given string (q-gram), locates all matches with

- Hamming distance ≤ 1 or edit distance ≤ 1
- rapidly
- without wasting memory
- with few false hits when hashing

Requires engineering appropriate hash functions.
Requires understanding statistics of hash functions.

Challenge

Technology-Dependent Read Mapping

- 454 & IonTorrent sequence a **run** of nucleotides at a time: TAAGTCCCA = (T, AA, G, T, CCC, A).
- Unable to determine exact length of a long run: AAAAAA $\approx$ AAAAAAA

Challenge

Technology-Dependent Read Mapping

- 454 & IonTorrent sequence a **run** of nucleotides at a time: TAAGTCCCA = (T, AA, G, T, CCC, A).
- Unable to determine exact length of a long run: AAAAAA $\approx$ AAAAAAA
- Idea: Ignore run length completely!
- Transform reference and reads by "forgetting": TAAGTCCCA = (T, AA, G, T, CCC, A) $\mapsto$ TAGTCA

Challenge

Technology-Dependent Read Mapping

- 454 & IonTorrent sequence a **run** of nucleotides at a time: TAAGTCCCA = (T, AA, G, T, CCC, A).
- Unable to determine exact length of a long run: AAAAAA $\approx$ AAAAAAA
- Idea: Ignore run length completely!
- Transform reference and reads by "forgetting": TAAGTCCCA = (T, AA, G, T, CCC, A) $\mapsto$ TAGTCA
- No two adjacent characters are equal.
- Build indexing / hash function on this property.
- Effective alphabet size: 3 instead of 4

Part 2

miRNA Expression in Neuroblastomas with SOLiD

miRNA Expression in Neuroblastoma (SOLiD)

SOLiD: short reads (35 bp), ideal for short non-coding RNAs
dinucleotide color space

1. read mapping
2. classification of reads
3. quantification of miRNA expression:
   normalization method
4. differential expression between neuroblastoma subtypes?
   detection of weak differential expression
5. miRNA-Editing?
6. discovery of two new miRNAs: now in miRbase

Schulte, . . . , SR, Schramm; *Nucleic Acids Research*, 2010.

Part 3

Determining CpG Island Methylation with 454

Bisulfite Sequencing of CpG Islands (454)

Goal

Determination of methlyation state in CpG islands

454-Technology: Pros und Cons

- relatively long reads
- compatible with bisulfite treatment (meth-C $\mapsto$ C, but C $\mapsto$ U=T)
- sequencing errors primiarily in *runs*, TTTTTT $\approx$ TTTTTTT
- problem: (close to) 3-letter alphabet, long runs

Read Mapping

In parallel against two genomes: bisulfite-treated, untreated Variant:
Shorten runs to one character in genomes and in reads

Library Optimization

Goal: Sequencing of CpG islands.
How to obtain them from the genome?

Library Optimization

Goal: Sequencing of CpG islands.
How to obtain them from the genome?
Restriction enzymes and length selection, but how to opimize them?

Library Optimization

Goal: Sequencing of CpG islands.

How to obtain them from the genome?

Restriction enzymes and length selection, but how to opimize them?

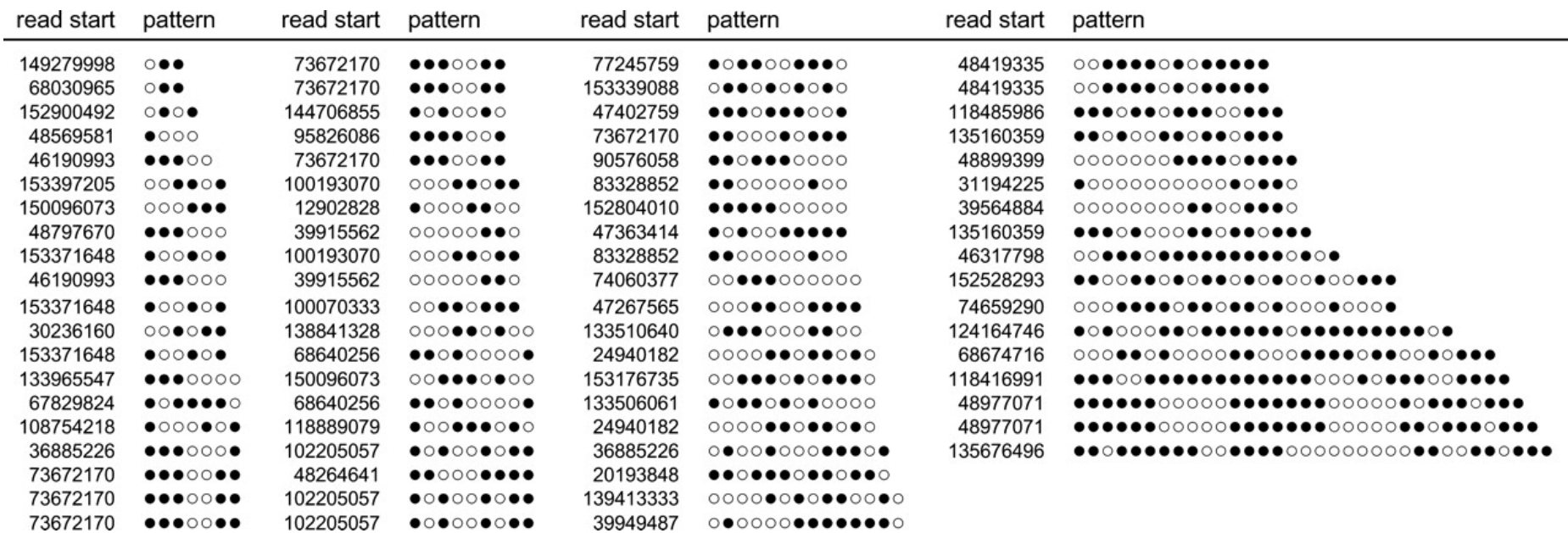
In-silico-optimization by simulation

Experiment Number	Enzyme Combination			Fragment Length Min	Fragment Length Max	Number of Distinct Fragments	CGI Score	Alu Score	#Fragments in MinLength -5% to +5%	#Fragments in MaxLength -5% to +5%	Fragments in [Min, Max] hitting a CGI	CGIs hit by Fragments in [Min, Max]
0	MseI	Tsp509		613	800	99388	5473	21472	53501	25524	4344	4035
1	MseI	Tsp509	AluI	436	800	99890	15652	16442	44360	3868	10785	9204
2	MseI	Tsp509	NlaIII	463	800	99529	18384	15019	44658	5808	13264	10778
3	MseI	Tsp509	BfaI	532	800	99980	11400	18664	43833	12026	8349	7416
4	MseI	Tsp509	HpyCH4	436	800	99207	18142	12440	43068	4189	12537	10417
5	MseI	Tsp509	DpuI	536	800	99748	11740	11799	44704	13026	8636	7623
6	MseI	Tsp509	MboII	573	800	99255	9250	20326	48504	17360	7040	6343
7	MseI	Tsp509	MlyI	588	800	99642	8597	17305	51587	19675	6544	5921
8	MseI	Tsp509	BCCI	574	800	99162	9387	18362	47642	17366	7158	6461
9	MseI	Tsp509	AluI	NlaIII	337	800	99463	22451	15126	44353	1364	14368	11217
10	MseI	Tsp509	AluI	BfaI	381	800	99403	17455	17371	45858	2067	11490	9539
11	MseI	Tsp509	AluI	HpyCH4	332	800	99268	19141	14561	47417	1178	11975	9672
12	MseI	Tsp509	AluI	DpuI	378	800	99736	16823	11089	43809	2129	11005	9223
13	MseI	Tsp509	AluI	MboII	400	800	99865	16930	17280	46676	2655	11283	9479
14	MseI	Tsp509	AluI	MlyI	412	800	99568	15229	15229	45205	2862	10340	8851
15	MseI	Tsp509	AluI	BCCI	401	800	99173	16790	16466	47211	2362	11203	9450
16	MseI	Tsp509	NlaIII	BfaI	398	800	99090	22430	16298	40045	2690	15153	11901
17	MseI	Tsp509	NlaIII	HpyCH4	346	800	99187	25121	9605	45792	1595	16312	12482
18	MseI	Tsp509	NlaIII	DpuI	402	800	99102	21570	10010	44201	3039	14649	11599
19	MseI	Tsp509	NlaIII	MboII	425	800	99924	20966	16022	43109	3828	14623	11713
20	MseI	Tsp509	NlaIII	MlyI	437	800	99464	19648	12975	42603	4127	13774	11134
21	MseI	Tsp509	NlaIII	BCCI	424	800	99988	21247	15236	43892	3688	14814	11731
22	MseI	Tsp509	BfaI	HpyCH4	372	800	98930	21213	13141	43407	2043	13845	11133
23	MseI	Tsp509	BfaI	DpuI	456	800	99422	15632	10735	40167	5825	10894	9294
24	MseI	Tsp509	BfaI	MboII	490	800	99888	14261	19726	42657	7703	10075	8700
25	MseI	Tsp509	BfaI	MlyI	502	800	99302	13143	16291	44338	8520	9413	8221
26	MseI	Tsp509	BfaI	BCCI	486	800	99301	14842	18137	43519	7224	10486	8988
27	MseI	Tsp509	HpyCH4	DpuI	378	800	99084	19885	8165	42925	2093	13052	10639
28	MseI	Tsp509	HpyCH4	MboII	397	800	99374	19939	13256	43682	2677	13309	10828

Patent with Roche Diagnostics/454

Result

Many CpG islands on the X-chromosome are incompletely methylated



Zeschnigk, . . . , SR, Horsthemke; *Human Molecular Genetics*, 2009

Part 4

Modeling and Finding Signals in Sequences

Probabilistic Methods for Signals in Sequences

$$\text{sequences} \xrightarrow[\text{probabilistic models}]{\text{statistical methods}\atop\text{efficient algorithms}} \text{signals}$$

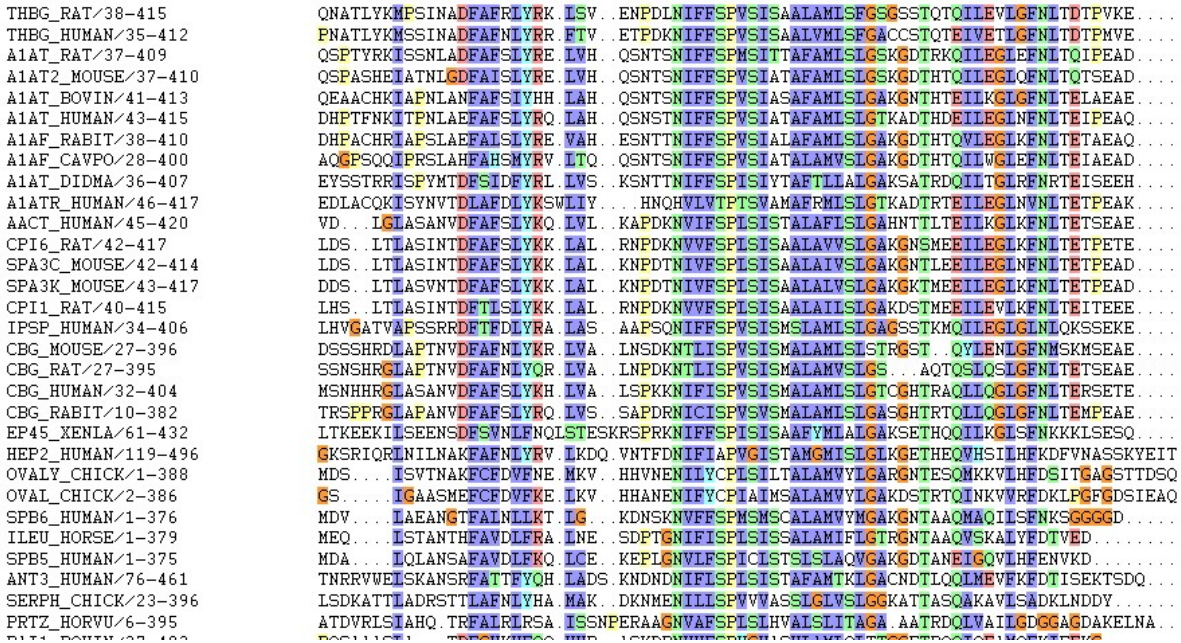
Examples for signals

- Repeats (exact, approximate)
- Overrepresented motifs
- evolutionarily conserved / variable regions
- SNPs, CNVs
- unique sequences (species identifications)
- core genome of a family

Protein families, HMMs and HMM-Logos

Pfam Database:

Multiple sequence alignments of representatives
from protein domains, folds, families

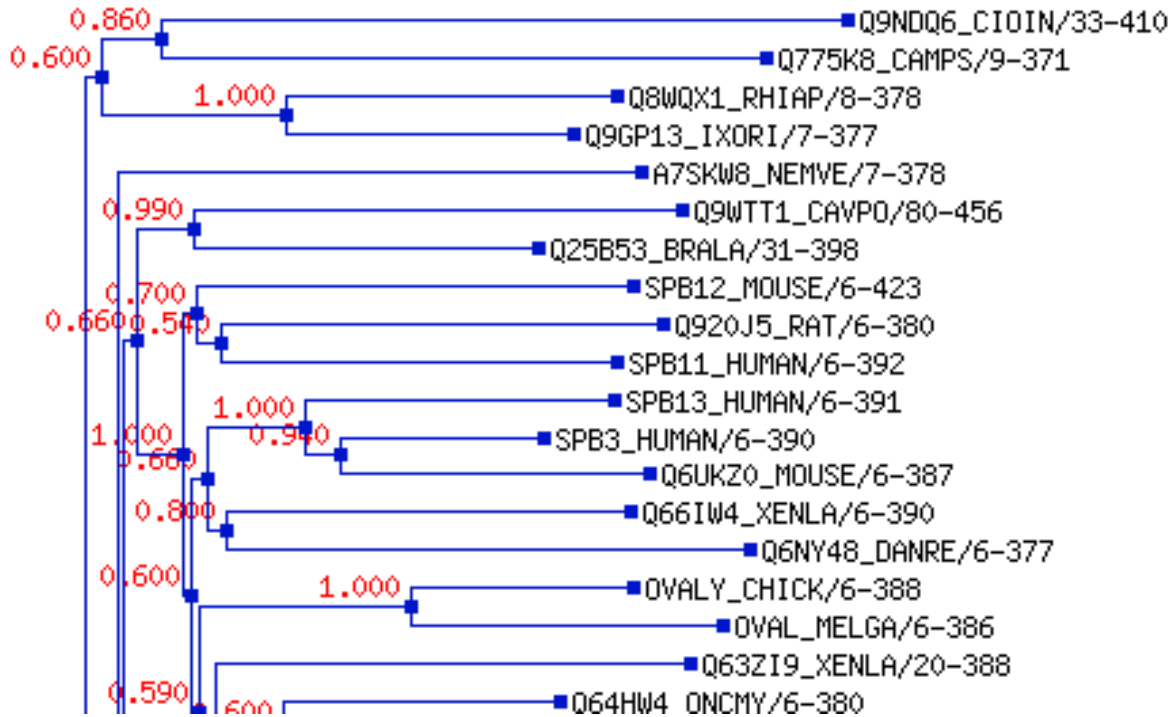


Alignment of serpins from Pfam

Descriptive, not very concise.
Horizontal and vertical perspective.

Horizontal Perspective: Gene Trees

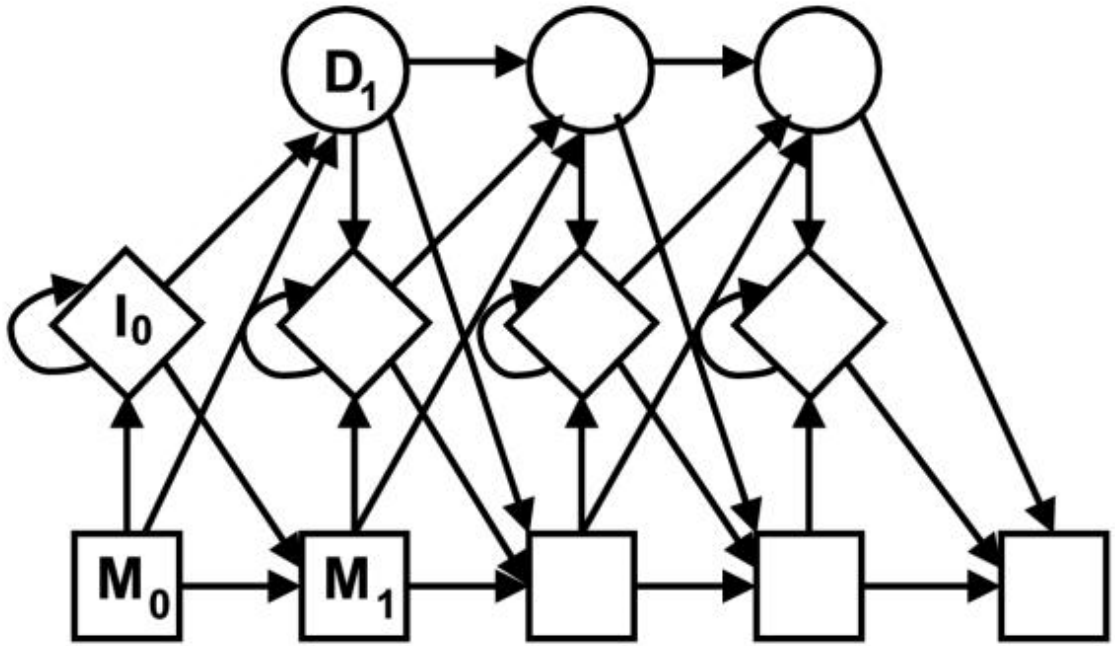
Objects of interest are sequences.
Distances, Similarities, Clustering.
Distances from Evolutionary Markov Processes
Gene Tree from clustering: (fast) UPGMA, (fast) Neighbor Joining



Serpin tree from Pfam

Vertical Perspective: HMMs

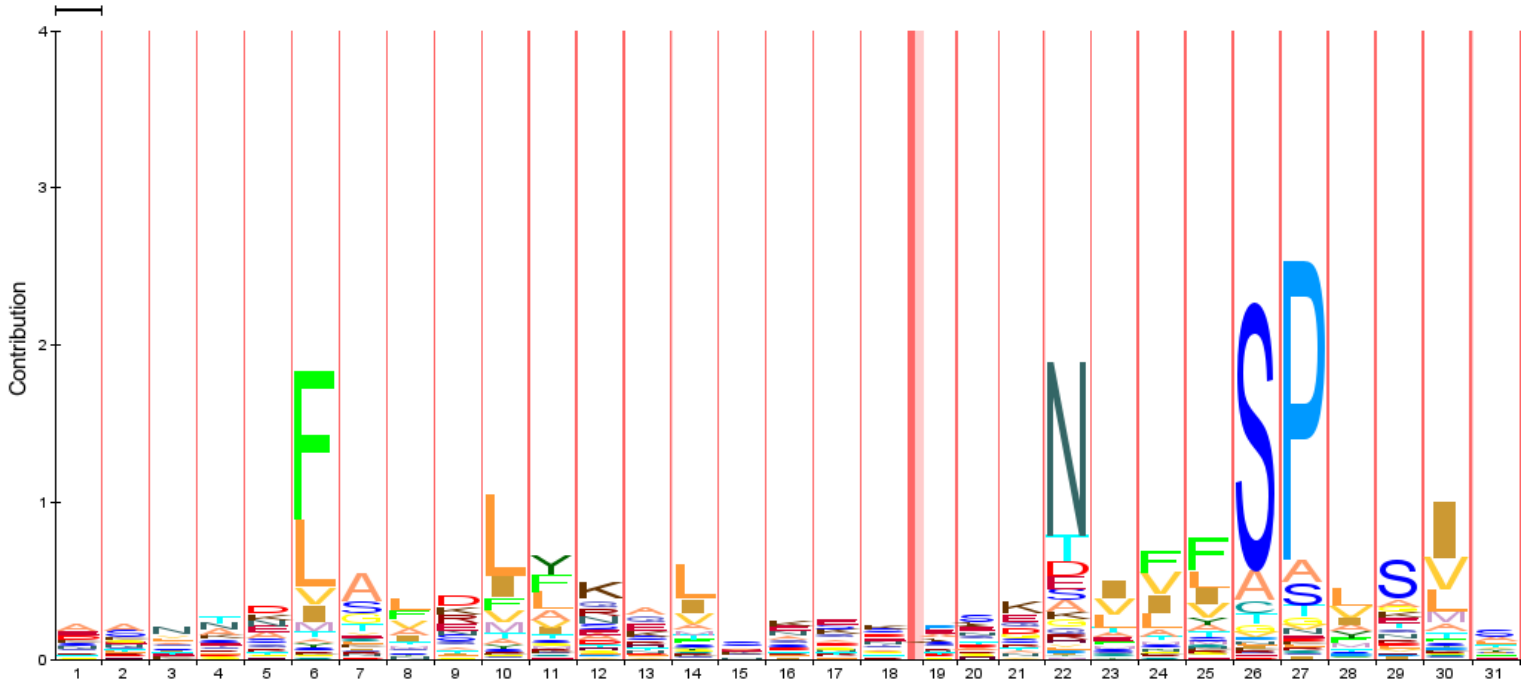
Objects of interest are positions (sites).
Conservedness, variability.
Possbily: Correlation with other positions.
Description by probabilistic model: HMM



Generative model, generalizes aligned sequences.

HMM constructed by HMMer, Eddy et al. (1994–2010)

Visualisation of HMMs by HMM Logos



Stack height	rel. entropy of position and background distribution
Symbol height	rel. frequency of amino acid
Stack width	1− deletion probability
Red bars	insertion probability and number of insertions

Does not represent horizontal correlations!

Schuster-Böckler, Schultz & SR (2004)

Species-Site Interactions

Usual views on multiple sequence aligments:

- horizontal: clustering, species trees
- vertical: probabilistic models (HMMs)

Species-Site Interactions

Usual views on multiple sequence aligments:

- horizontal: clustering, species trees
- vertical: probabilistic models (HMMs)

Joint view delivers more information:

- Which sites are responsible for which splits in the tree?
- Are there signs of recombination?

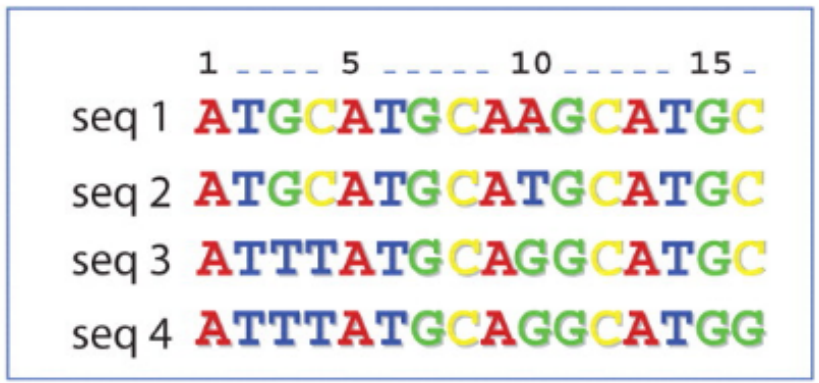
Species-Site Interactions

Usual views on multiple sequence aligments:

- horizontal: clustering, species trees
- vertical: probabilistic models (HMMs)

Joint view delivers more information:

- Which sites are responsible for which splits in the tree?
- Are there signs of recombination?



Explorative Analysis: Ideas

1. Embed the sequences into in $\mathbb{R}^n$
2. Visualize the embedded data in $\mathbb{R}^2$

Explorative Analysis: Ideas

1. Embed the sequences into in $\mathbb{R}^n$
2. Visualize the embedded data in $\mathbb{R}^2$

Embedding

- Fisher Scores from an HMM of the whole alignment
  Measure of influence of sequence S_i on HMM parameter θ_j:

$$f_{ij} = \nabla_{\theta_j} \log \mathbb{P}[S_i \mid \theta]$$

- Properties:
  - Encodes emission, deletion, and insertion probabilities
  - Efficiently computed (with a Forward-Backward-type algorithm)
  - HMM model allows to incorporate external knowledge
  - HMM parameters directly would not depend on sequences.
  - Disadvantage: High-dimensional representation

Correspondence Analysis

Pre-Processing

Normalized data matrix H from Fisher-Matrix F:

$$H := R^{-1/2} \cdot (F + \sigma) \cdot C^{-1/2}$$

Singular Value Decomposition

$$H = U \Sigma V^T,$$

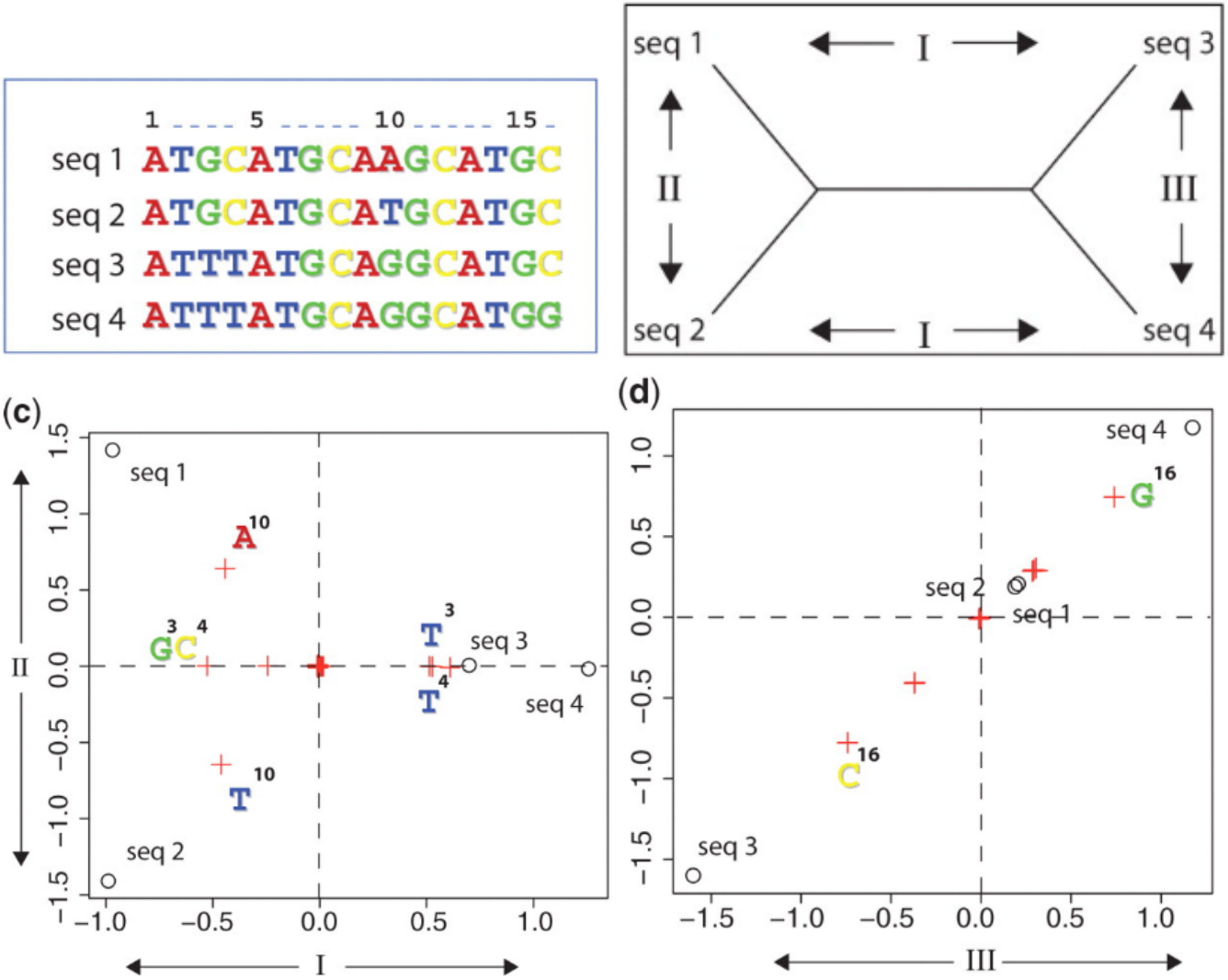
where U, V orthogonal, $\Sigma \geq 0$ diagonal.
Σ: "Singular values", ordered decreasingly.
U, V: left resp. right singular vectors.

Post-Processing

Rescaling of U, V: Scaled u_i, v_i = Principal Axes.
Plotted into joint coordinate system.

Schwarz, Seibel, SR et al., *Nucleic Acids Research*, 2009

Example



Left: First and second principal axes. Right: Third principal axis.

Applications (Univ. Würzburg)

- *Neisseria meningitidis*,
  Factor H binding protein (fHBP) = lipoprotein LP2086
  114 sequences (47 distinct ones)
  Alignment gives conflicting signals.
- Vitamin K Epoxid Reductase (VKORC1), paralog VKORC1L1

Schwarz, Seibel, SR et al., *Nucleic Acids Research*, 2009

Part 5

The Future?

Challenge: Probabilistic Genome Models

Situation around 2000

Human genome almost done. Nothing left do do...

Challenge: Probabilistic Genome Models

Situation around 2000

Human genome almost done. Nothing left do do...

Current "Genome" Projects

- 1000-Genomes-Project (human pangenome): over 3 Tbp sequences
- International Cancer Genome Consortium
- Human Gut Metagenome Initiative (100 bacteria per human cell, gene pool 100x bigger)

Image: M. Gerstenberg
Die ZEIT (12/2006)

Pangenome := entirety of genetic information of a species
Metagenome := $\sim$ of a community

Future Challenges

Sample questions to a pangenome

- What's the genome of the 334-th sequenced person?
- How often and where does the motif TATAAW occur?
- Which variants of the dopamine D2 receptor gene exist?
- Which variables do these variants correlate with?

Which data structures provide this information?

- lossless sequence representation
- fast search (index based), also approximate
- representation of consensus and variations (e.g., SNPs, CNVs)
- representation of rearrangements, repeats
- generalization ability
- integration of annotation and semantics

Future Challenges

Sample questions to a pangenome

- What's the genome of the 334-th sequenced person?
- How often and where does the motif TATAAW occur?
- Which variants of the dopamine D2 receptor gene exist?
- Which variables do these variants correlate with?

Which data structures provide this information?

- lossless sequence representation
- fast search (index based), also approximate
- representation of consensus and variations (e.g., SNPs, CNVs)
- representation of rearrangements, repeats
- generalization ability
- integration of annotation and semantics

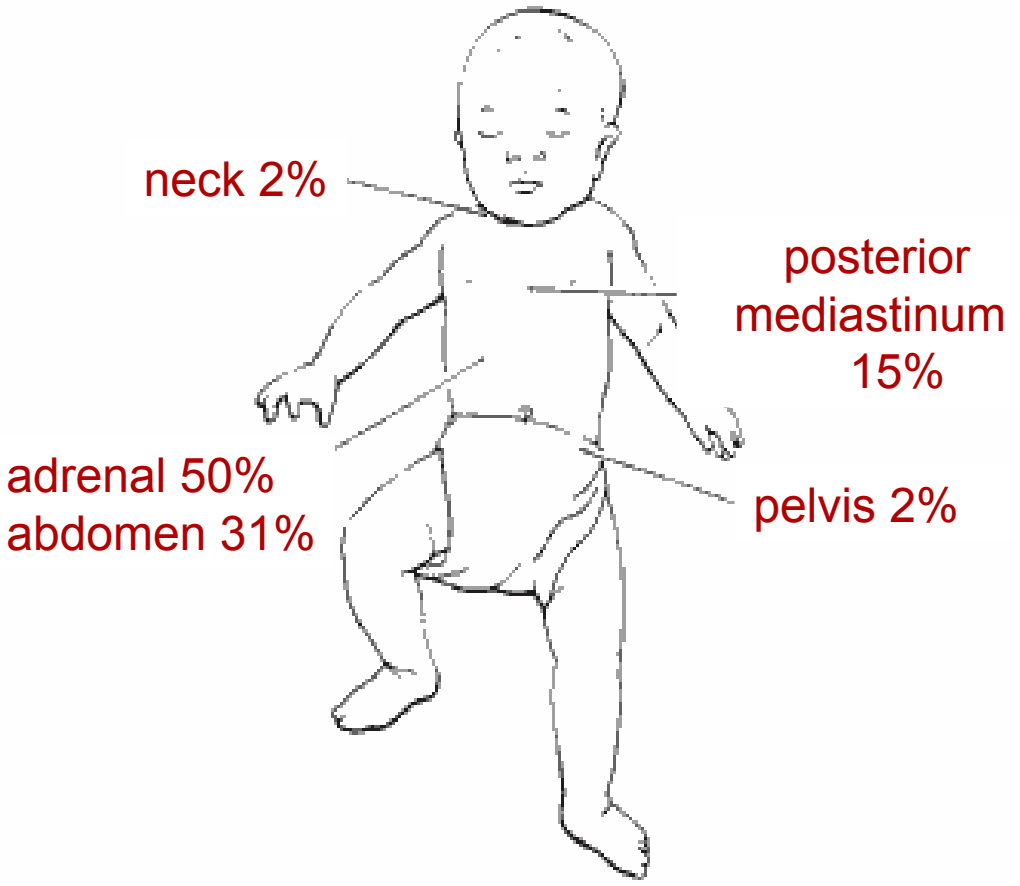
No existing ones.

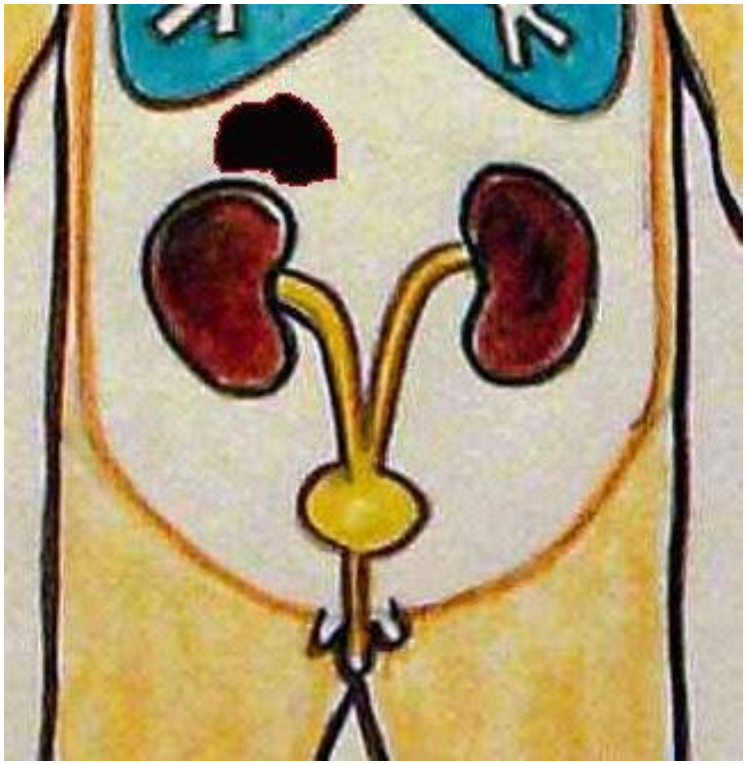
Neuroblastoma: Frequency and Origin

- Most frequent extracranial solid tumor of childhood (8-10%, 1 in 7000 births)
- 15% of all childhood cancer deaths
- Poor prognosis of high-risk neuroblastoma

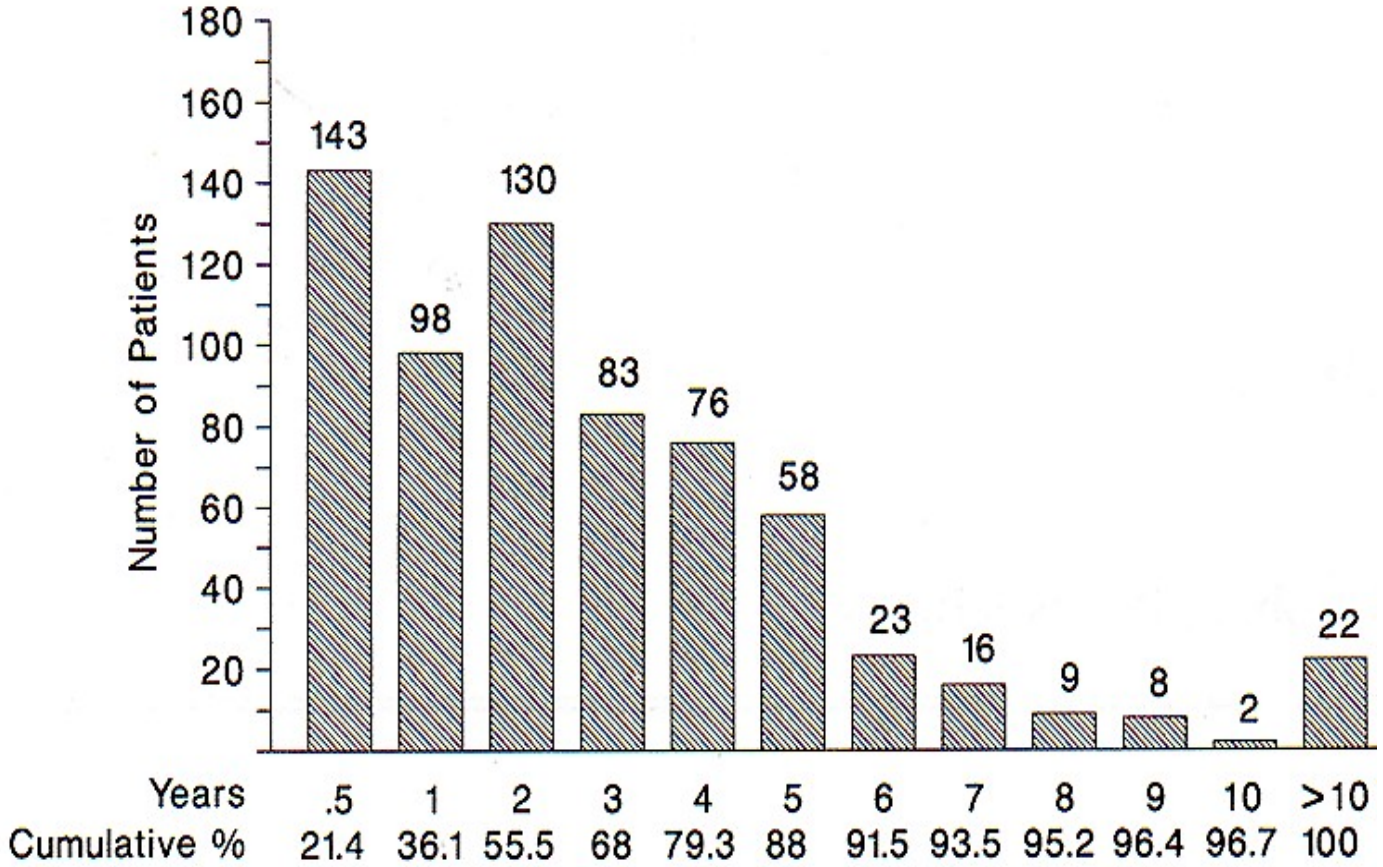
- Origin: postganglionic sympathetic neuroblasts (neural crest progenitor cells)
- Localisation: adrenal glands and cross chain

Neuroblastoma: Localisaton



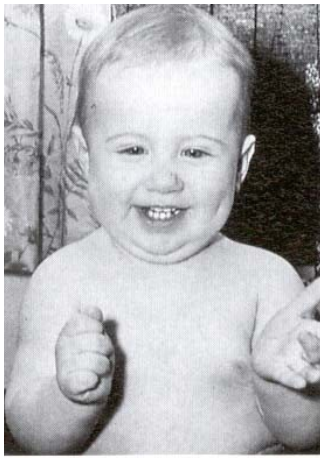


Neuroblastoma: Patient Age at Diagnosis



Neuroblastoma: Present Situation

- Clinical heterogeneiety: favorable vs unfavorable neuroblastoma



Stage I
good prognosis
favorable biology

Stage IV
bad prognosis
unfavorable biology

- Stage IVs: Spontaneous regression of metastasised disease
  - Differentiation to benign Ganglioneuroma

- Genetic model exists, but is incomplete.
- Transcriptome and proteome have been analyzed.
- New sequencing technologies will allow deeper insight.

Deep small RNA Transcriptome Sequencing: Motivation

- Unbiased, unselected identification of transcripts
- Absolute and exact quantification, good dynamic range
- Analysis of transcript sequence, including sequence variants by
  - SNPs
  - RNA editing
- Strand-specific expression analysis
  - miRNA-5p vs. -3p
  - miRNA* vs. miRNA

Study Cohort:
5 Favorable vs 5 Unfavorable Neuroblastoma

Pat. No.	Stage	Age at Dx	MNA	DoD	EFS	OS
552	1	405	0	0	3109	3109
553	1	481	0	0	3745	3745
554	1	961	0	0	3605	3605
555	1	459	0	0	2861	2861
556	1	103	0	0	2856	2856
557	4	1478	1	1	946	1375
558	4	496	1	1	351	539
559	4	1045	1	1	839	1115
560	4	978	1	1	184	212
561	4	4827	1	1	201	207

- Age at Dx: Age at diagnosis
- MNA, DoD: MYCN amplified?, Died of Disease?
- EFS, OS: Days of event-free (resp. overall) survival

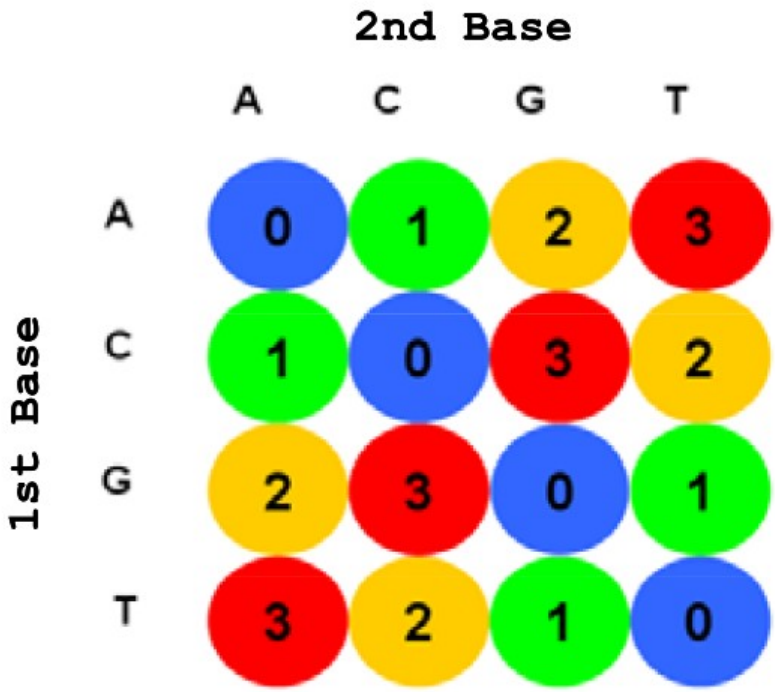
Sequencing Technology

- ABI SOLiD sequencing (35nt reads, color space)
- Small RNA Expression Kit (SREK)

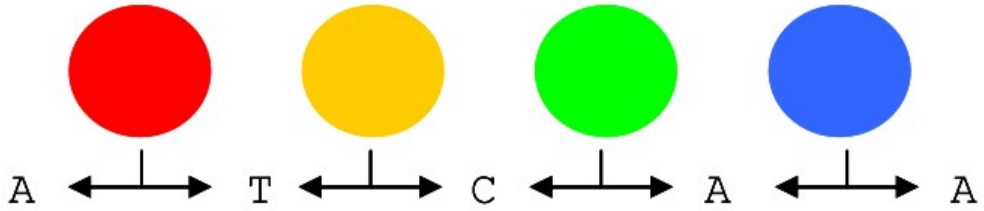
- SREK sequencing with ABI SOLiD of 10 patients
  - 10 separate ‚fields' of 1 slide
  - 188,821,076 reads total
  - number of reads varied widely by patient

SOLiD Color Space

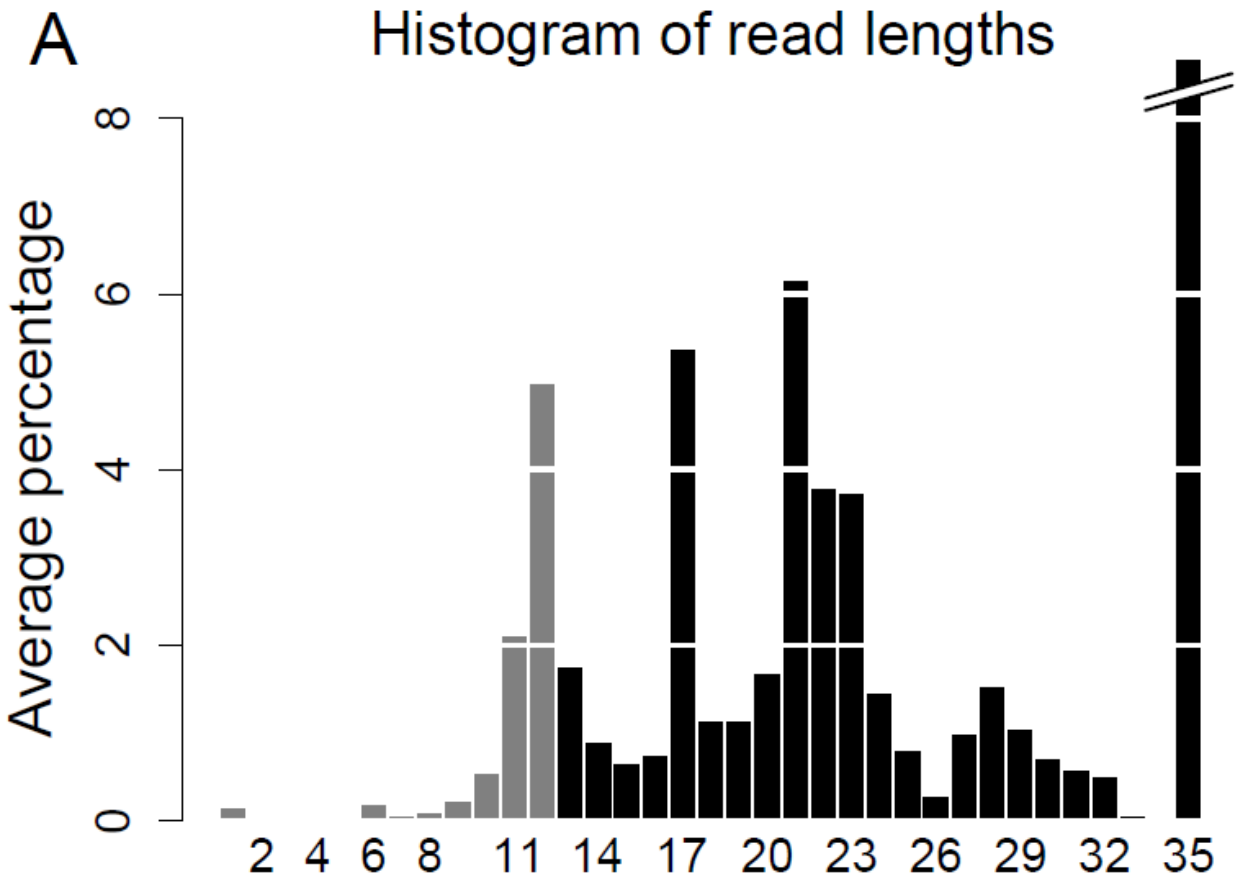
- ATCAA:
  A3210
- Each color
  represents
  a dinucleotide
- When mapping
  against a reference,
  this can be used
  for error correction



Double Interrogation: Each base is defined twice



Sanity Check: Distribution of Read Lengths (nucleotide space, after adapter removal)



removed: < 13 nt

microRNAs: 21-22 nt

full length (35 nt): 57%

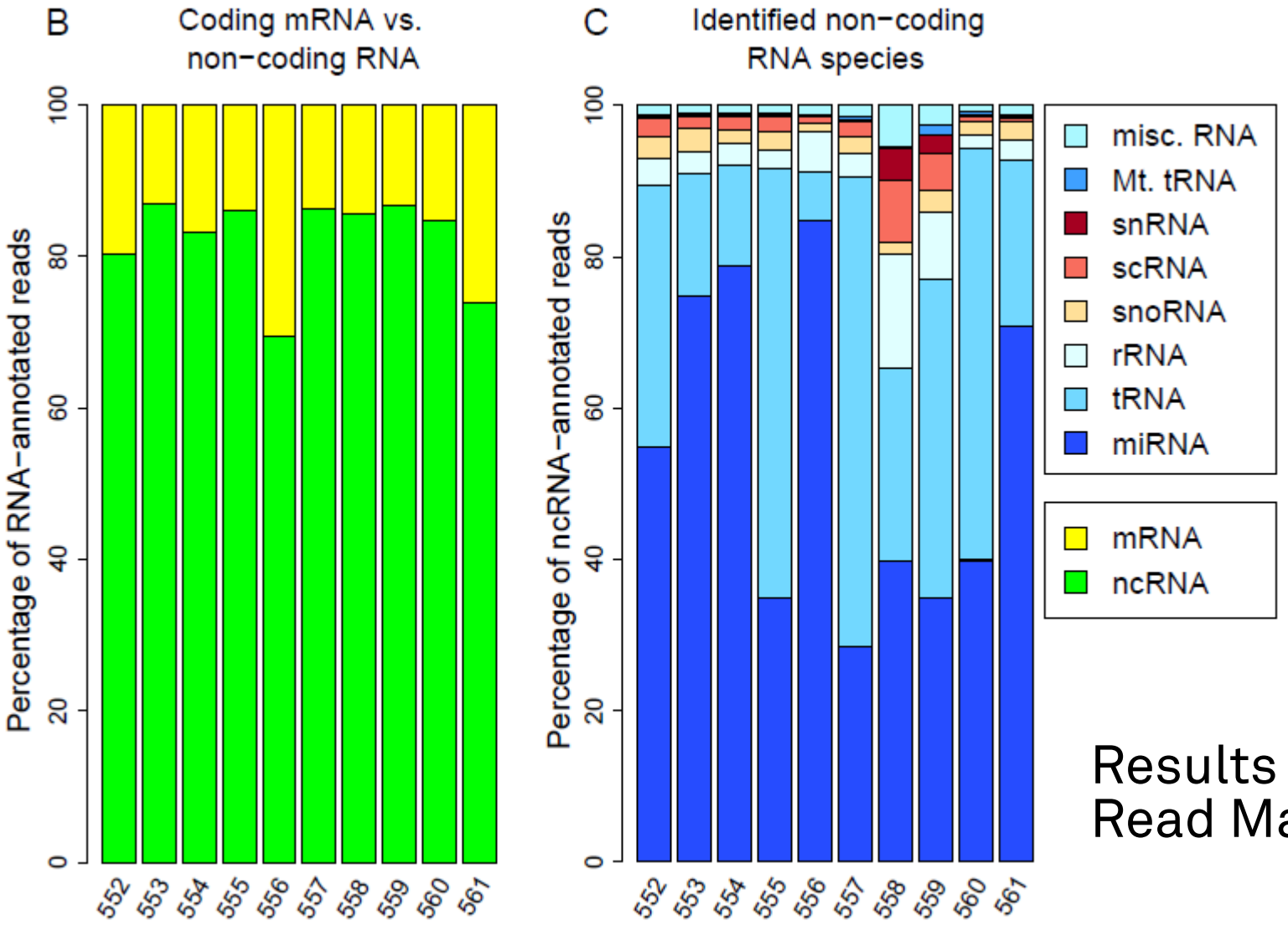
Method: Adapter removal

- Reads obtained in SOLiD dinucleotide color space (35 col.)
- Expected length of mature miRNAs: 20–23 nt
- miRNA reads contain part of adapter 330201030313112312
- Custom software (free-end-gap / semiglobal alignment with <12% errors) used to locate start of adapter
- Computation of full alignments requires quadratic time, but adapter and reads are short: time < 1 minute / million reads.

```
adapter sequence:                          330201030313112312
original read: T3000232100101222222330201030313112
trimmed read:    00023210010122222
```

Methods for Read Mapping

- Reads in color space, available at NCBI acc. no. SRA009986.
- References in nucleotide space:
  - Human Genome RefSeq Hg18
  - miRBase release 13.0
  - fRNAdb v3.1
  - RepBase 14.06
  - Human UniGene sequences (July 2009)
  - E. coli (NCBI Nucleotides accession no. NC_000913).
- Mapper: MAQ 0.7.1-10 (Subversion rev. 687): ungapped alignment only
- Work around bug in MAQ for short reference sequences (add flanking Ns)
- Allow <=2 errors (read length 12-14), or <= 3 errors (read length >= 15)
- Discard all non-unique matches (conservative approach)
- Convert mapped reads from color to nucleotide space, MAQ csmap2nt



Results of Read Mapping

Methods: Determining miRNA Expression Levels - Normalization

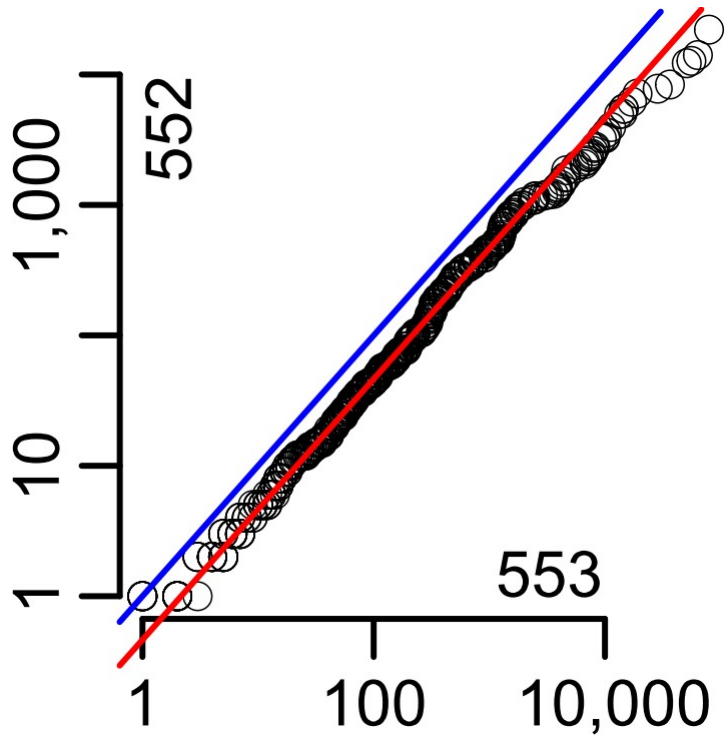
- Total number of reads varied widely by patient.
- Same for reads uniquely mapped to miRNAs.
- Normalization required.

How to normalize?

- Cannot apply microarray methods (only 465 miRNAs, not 1000s of genes!)
- Don't normalize (take raw data)
- Divide by sum (assumes total expression is equal)
- Robust linear transformation based on qq-plot

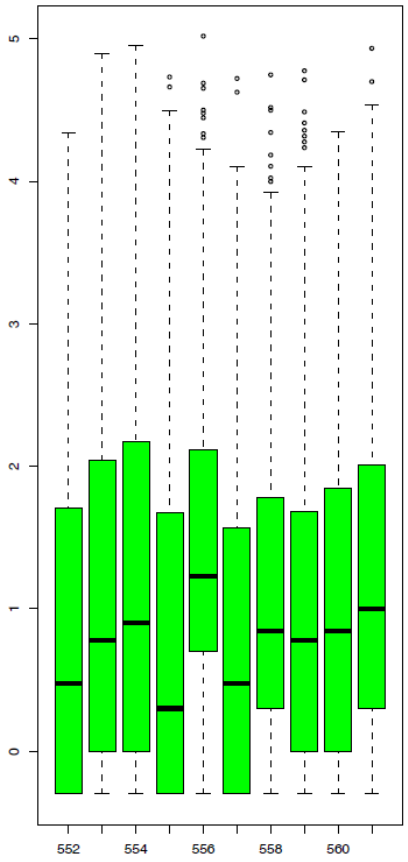
qq-Normalization

- Choose one dataset as reference (here: 552)
- Consider qq-plot between each dataset (example: 553) and reference
- Robustly fit a line through qq-plot points in log-space (red line)
  - minimize median of differences of log-expressions
- If inclination = 1 (log-space),
  determine shift to main diagonal (blue)
- Add shift to each expression value
  - corresponds to pure scaling transformation

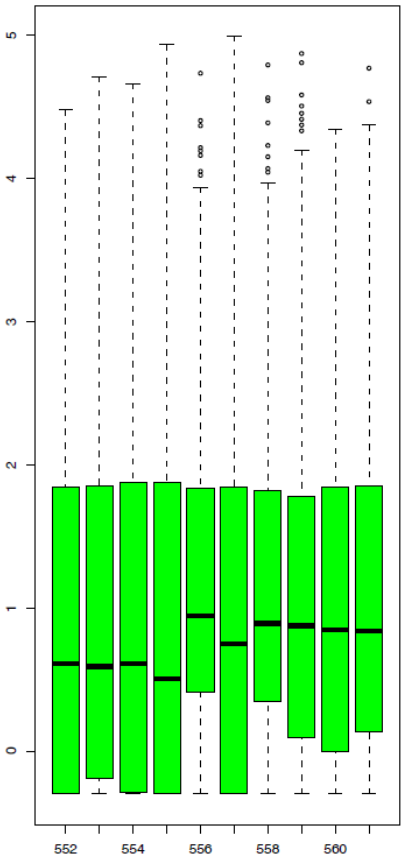


Comparison of Different Normalization Methods

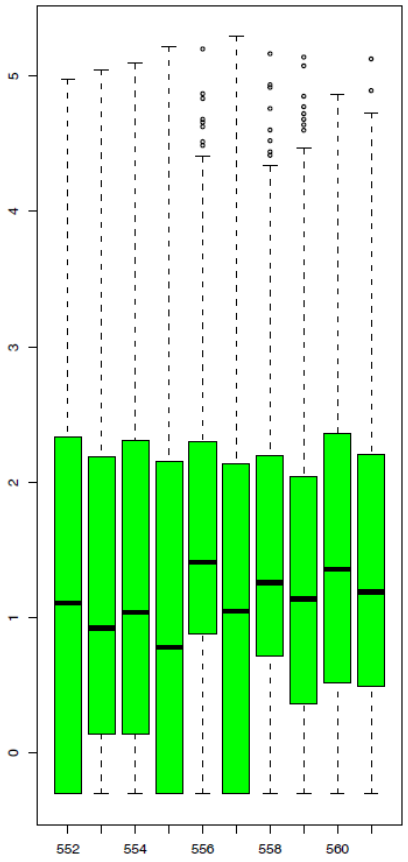
Raw Data



qq-Normalization



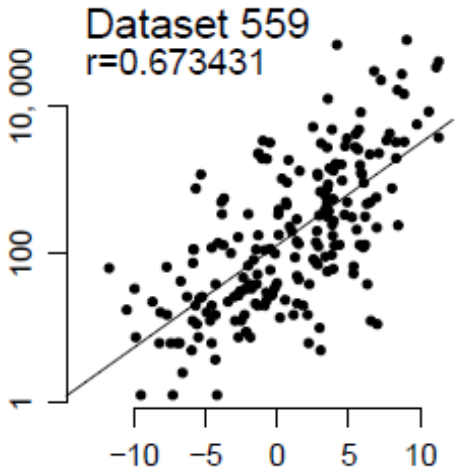
Divide by sum

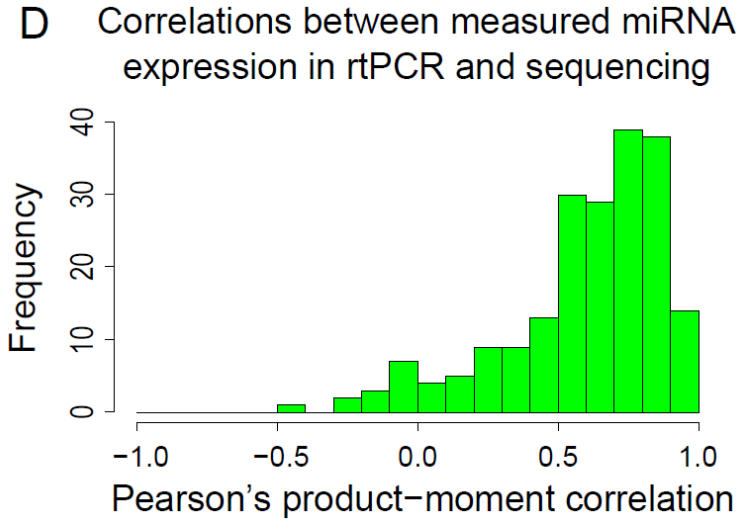


Boxplots show log-expression of 465 miRNAs for each patient

Validation: Comparison with RT-qPCR

- Left: one dataset (559), scatterplot
  normalized log-expression vs. RT-qPCR (-C_t value)
  Pearson correlation 0.67
- Right: Each of 204 evaluated miRNAs yields one correlation value
  over the 10 patients, normalized log-expression vs. RT-qPCR (-C_t value)

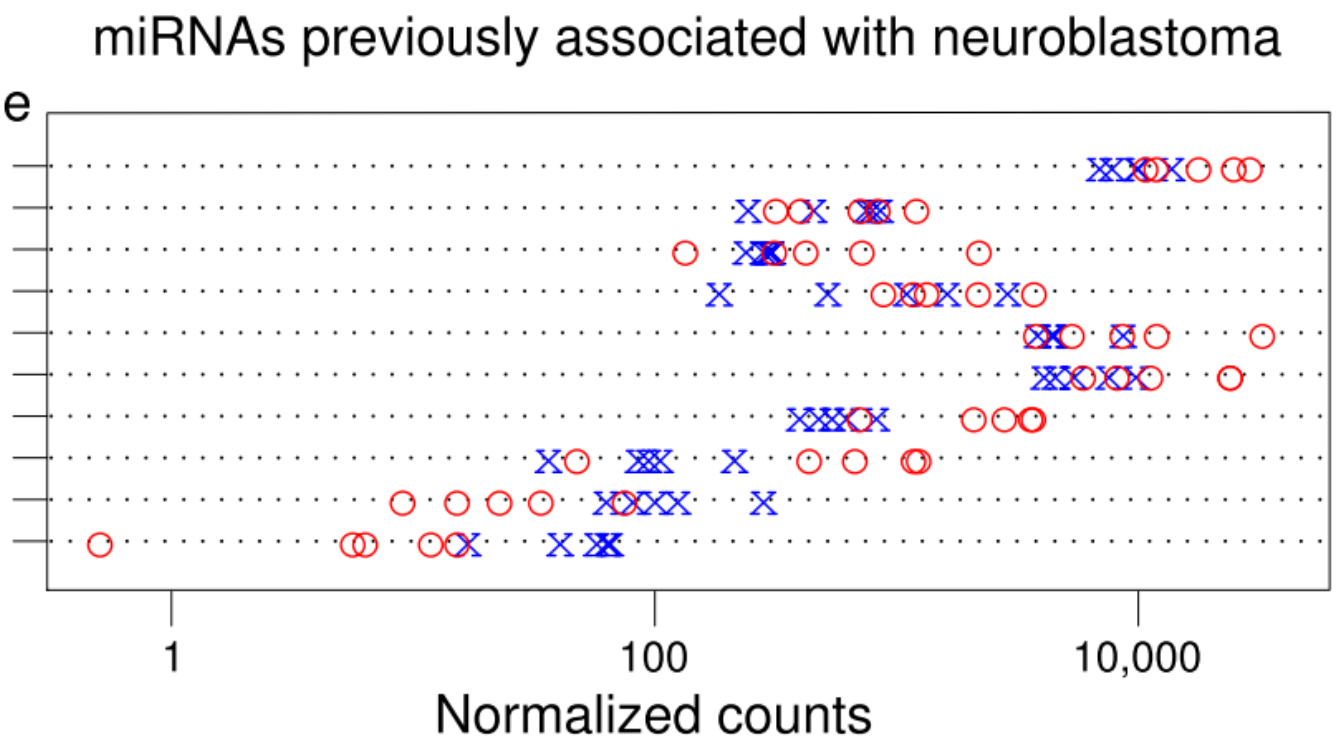




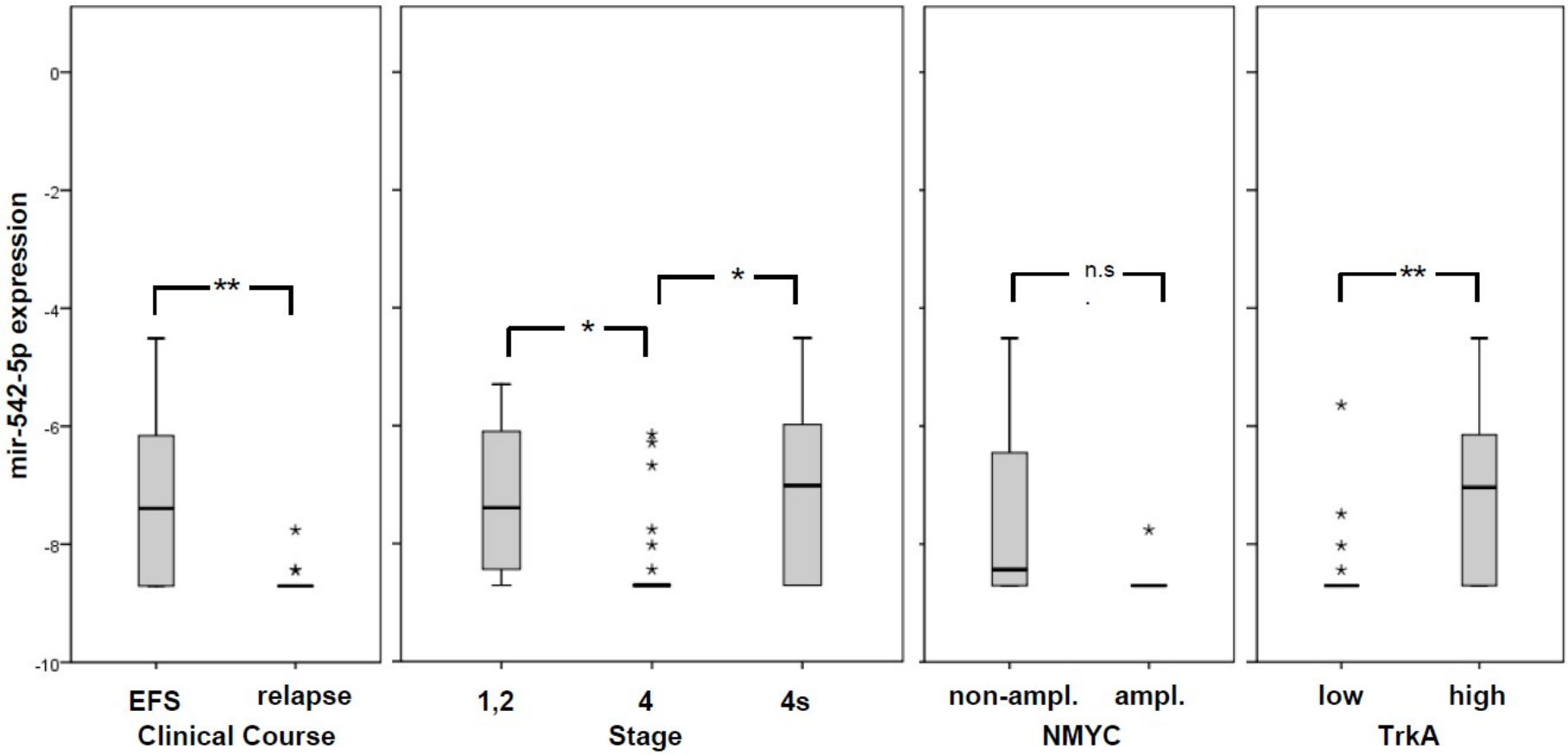
Expression of miRNAs previously reported as relevant to NB: Significant differential expression is hard to find.

A

miR	raw p–value	adj. p–value
17	0.0292	0.0730
18a	0.7563	0.7563
19a	0.2961	0.3290
20a	0.2817	0.3290
19b	0.1663	0.2375
92a	0.0557	0.0929
181a	0.0071	0.0493
181b	0.0518	0.0929
542–5p	0.0099	0.0493
628–5p	0.0197	0.0656

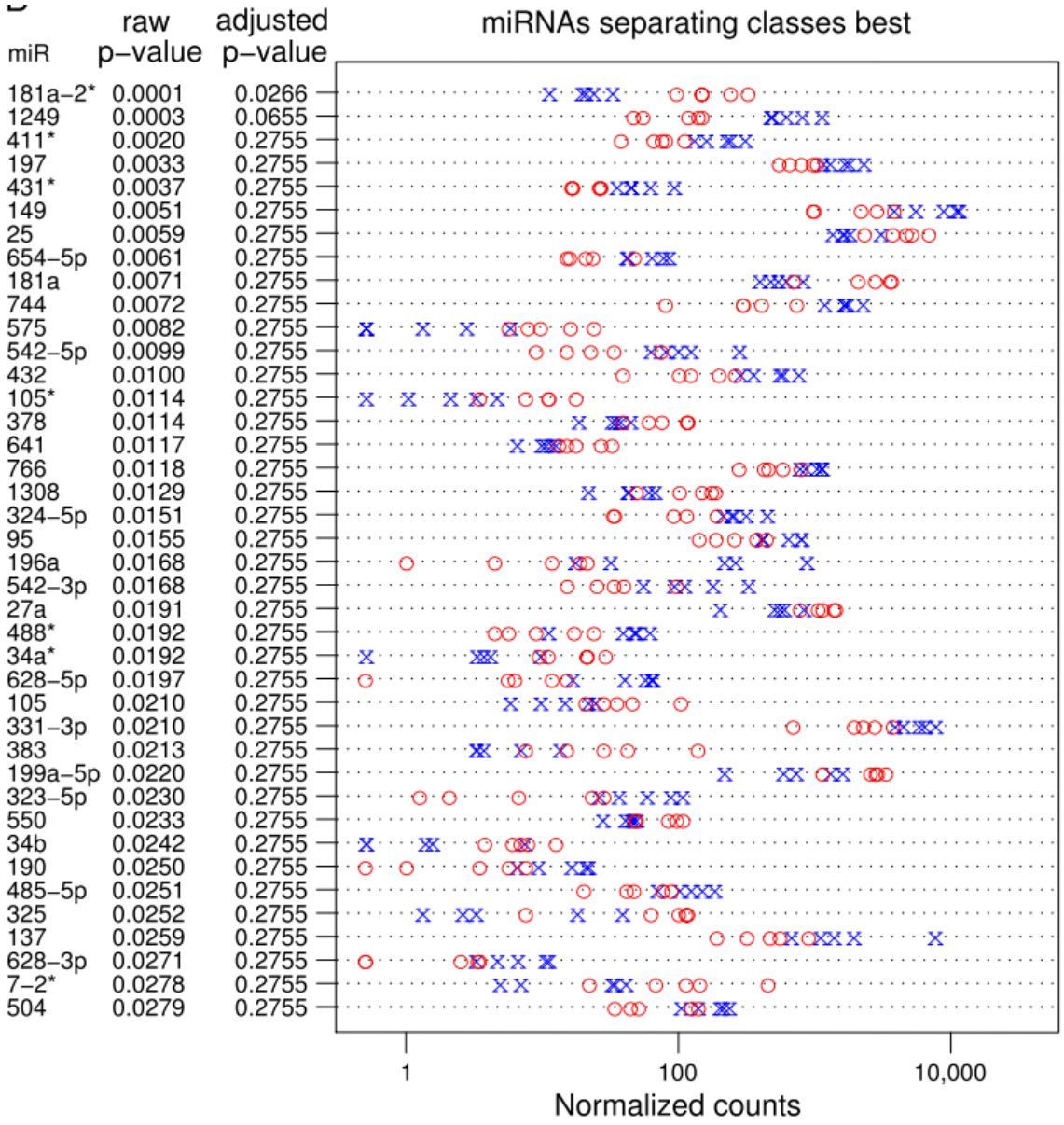


Validation of differential expression of miR-542-5p in 69 patients with RT-qPCR



Expression of Top 40 separating miRNAs

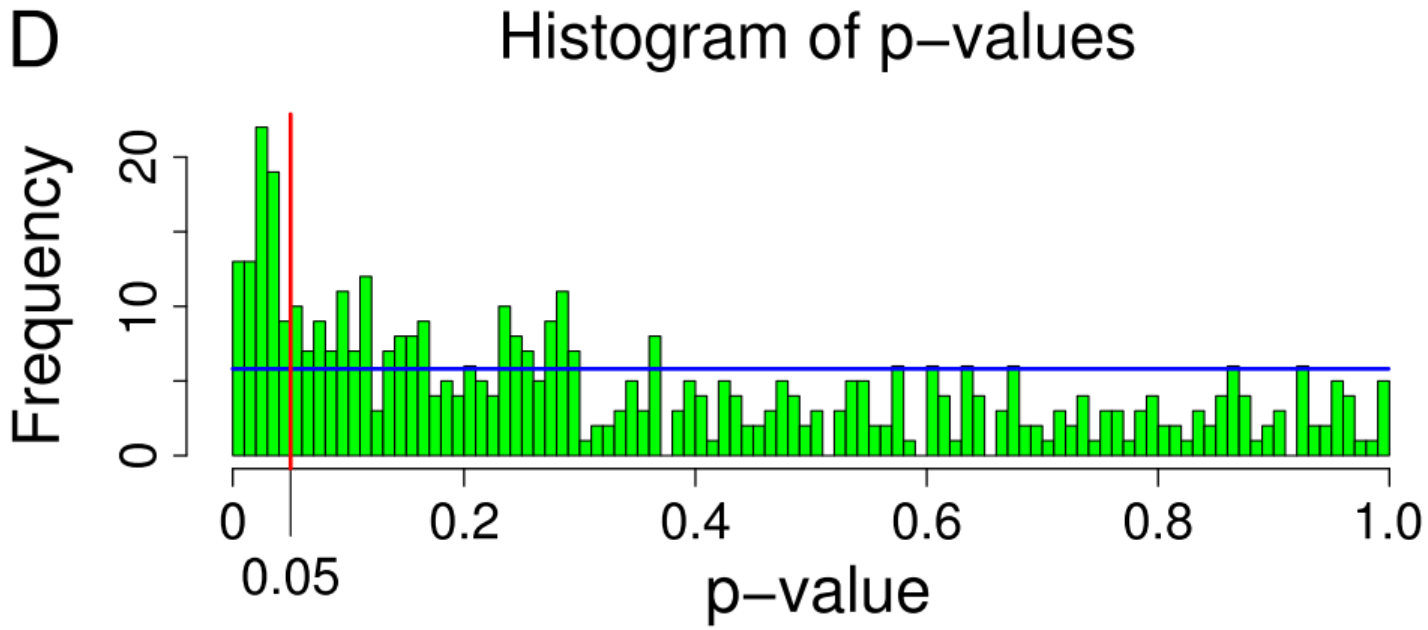
- several significant class-separating miRNAs before FDR correction
- only one significant class-separating miRNA after FDR correction for 465 tested miRNAs (Benjamini-Hochberg)



- unfavorable NB
- favorable NB

A Global View on Differential Expression: Distribution of Raw p-Values

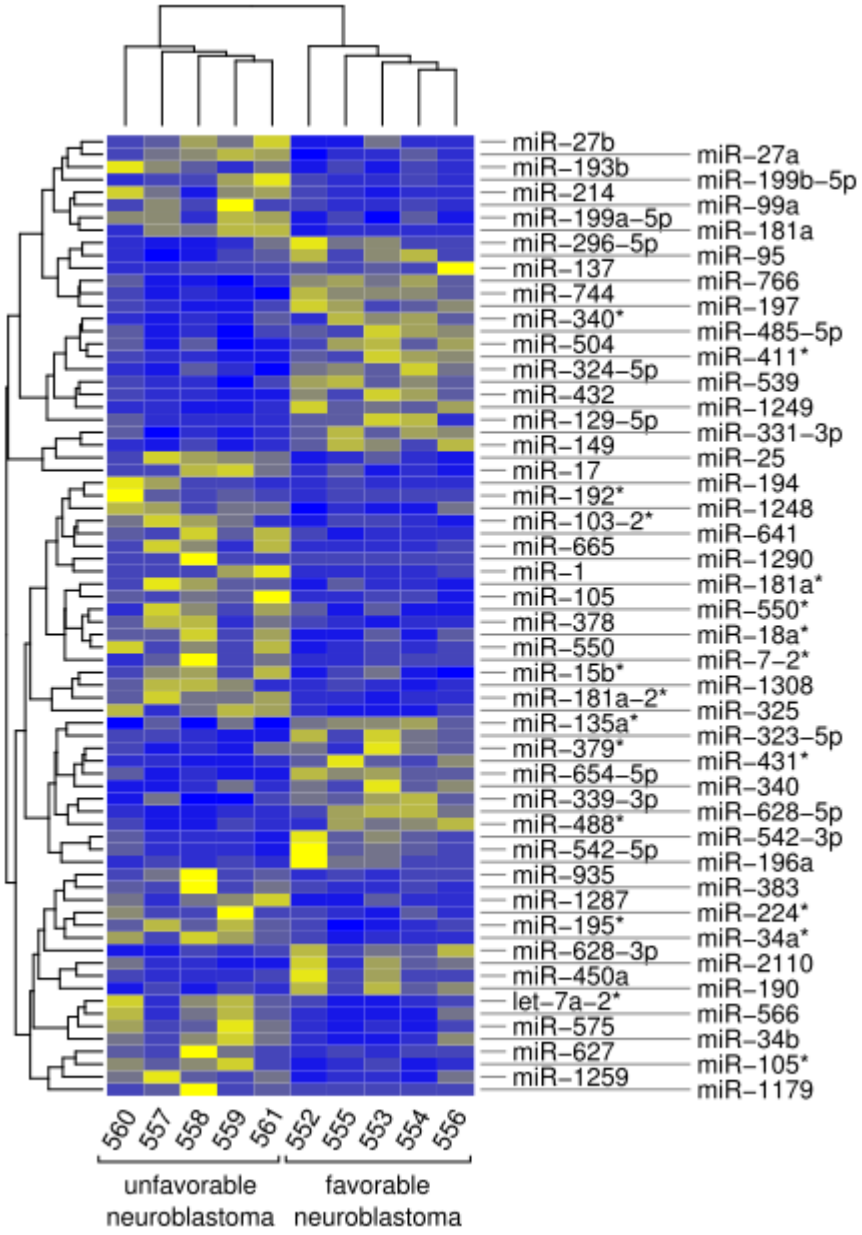
- Non-uniform distribution of 465 raw p-values:
  significant global differential expression,
  only few transcripts reach significance after multiple testing correction
  (small sample size)



Clustering: Perfect Separation

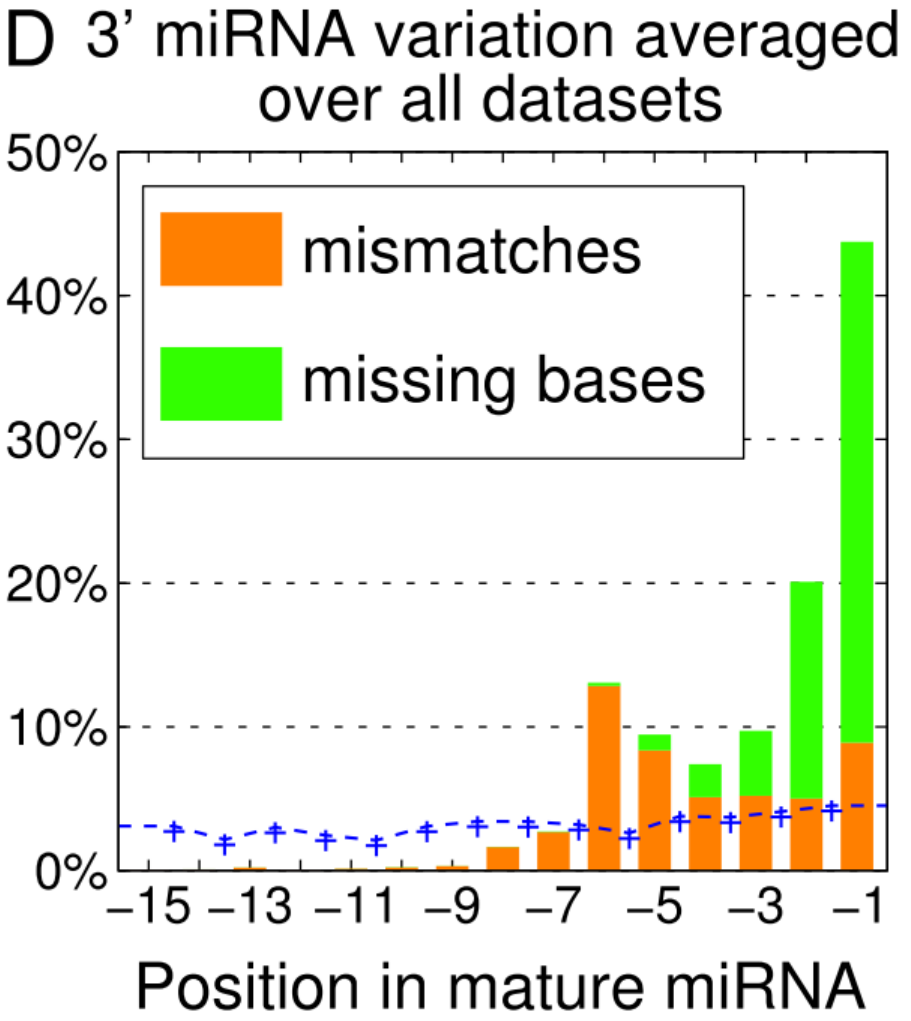
- 76 significant miRNAs (uncorrected)
- Hierarchical single-linkage clustering using 'heatmap' (R 2.9.1)
- Canberra Distance on normalized expression values

- Same perfect separation when all 465 miRNAs are used.



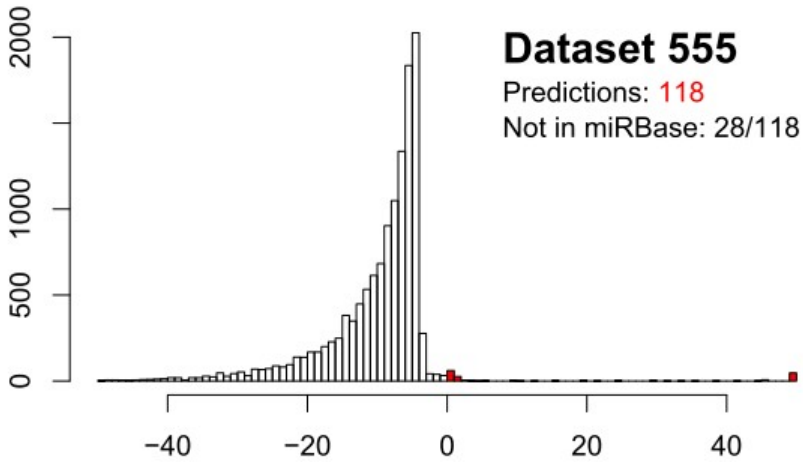
Global View on Editing

- Significant editing at positions -6, -5 from 3' end.
- Blue line / crosses: estimated sequencing error probability in color space. Lower in nucleotide space, as errors can be corrected.



Methods: De-novo discovery of putative new miRNAs

- Tool: miRDeep software from Max-Delbrück Center, Berlin
- Custom re-implementation of script `excise_candidates.pl` with different efficient data structures, avoiding quadratic time behavior.
- RNAfold from Vienna package (v.1.8.2) to predict structure
- miRDeep score histograms were consistent with published references:



Score >1 indicates good prediction.
Positive scores should be rare.

- with blastclust (BLAST v.2.2.20)

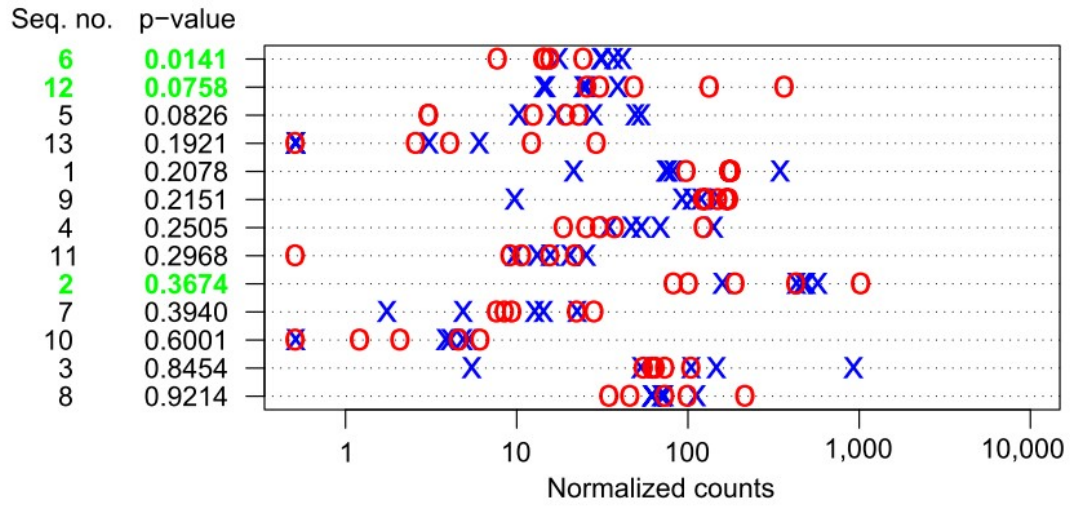
miRDeep Results

- 64% of predicted miRNAs exactly matched an entry in miRBase
- 24 sequences contained no known miRNA motifs,
  and were represented in at least three different datasets.
- 13 of these 24 had no BLAST similarity (E>0.1) to known miRNA sequences.
- 13 strong candidates for novel miRNAs

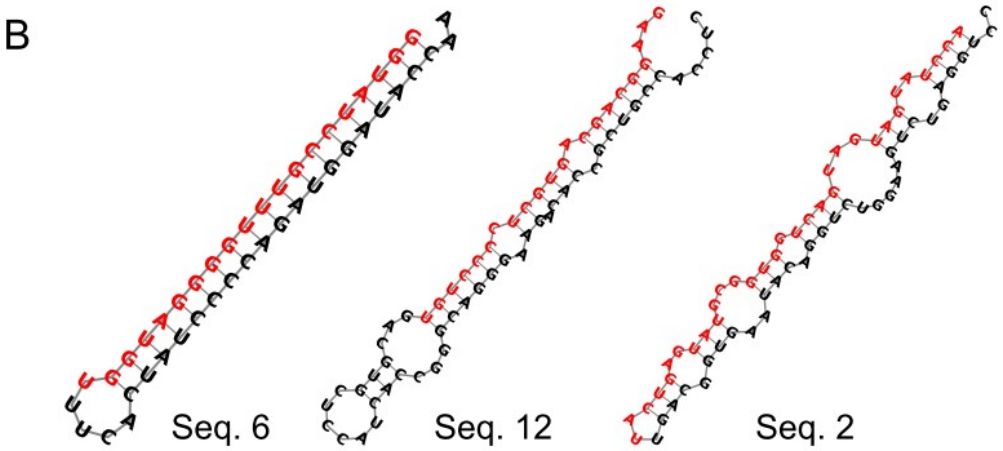
- 3 miRNAs selected for validation and validated with RT-qPCR
  - Seq 6, Seq 12: differential expression
  - Seq 2: high expression
- Secondary structure prediction: expected stem-loop configuration.
- RT-qPCR confirmed expression of Seq 2 in 69 out of 70 primary NBs .

Results: De-novo discovery of putative new miRNAs

- Expression values of 13 discovered miRNAs



- structure of 3 validated discovered miRNAs



Summary

- Next Generation Sequencing (NGS):
  - new tool to adress the complexity of small RNA transcriptomes
  - reveals insights into the miRNA world
- Pilot study to compare small RNAs of 5 favorable vs. 5 unfavorable NB.
- Unbiased, absolute quantification of small-RNA transcriptome with NGS.
- Normalization is an issue.
- High correlation of normalized NGS with stem-loop RT-qPCR data.
- Globally differential miRNA expression observed.
- Putative tumor suppressive miR-542-5p differentially expressed.
- Extensive miRNA editing:
  No systematic difference but individual miRNAs differentially edited
- 13 new putative miRNAs identified by modified miRDeep algorithm

Future Bioinformatics Challenges for Small RNA Sequencing

- Discuss Normalization
- Cross-Mapping
  - Some small RNAs have similar sequences.
  - One highly expressed RNA will generate some erroneous reads that may register es perfect reads from a different unexpressed RNA.
  - Sequencing errors vs. editing vs. SNPs vs. similar RNAs
  - Color space error correction may have saved us some trouble.
- Analysis Pipeline Improvements
  - At the moment: Considerable manual work
  - Combination of Python and R scripts, (somewhat buggy) external tools (MAQ, miRdeep)