GENE EXPRESSION PROFILING BY MASSIVELY PARALLEL SEQUENCING

by Tatiana Teixeira Torres, Muralidhar Metta, Birgit Ottenwlder and Christian Schltterer

Presenter: Ruslan Salakhutdinov

March 26, 2008

Paper in a nutshell

- The goal of this paper is to evaluate the potential of 454 sequencing technology to serve as a reliable tool for expression profiling.
- The result of the study is that using random breakage of the cDNAs by nebulization, 454 sequencing can be successfully used for expression profiling.
- Here, complementary DNA (cDNA) is DNA synthesized from a mature messenger RNA.
- The sequenced fragments can be mapped with high accuracy onto the Drosophila melanogaster genome.

Gene Expression Profiling

Source: Wikipedia.

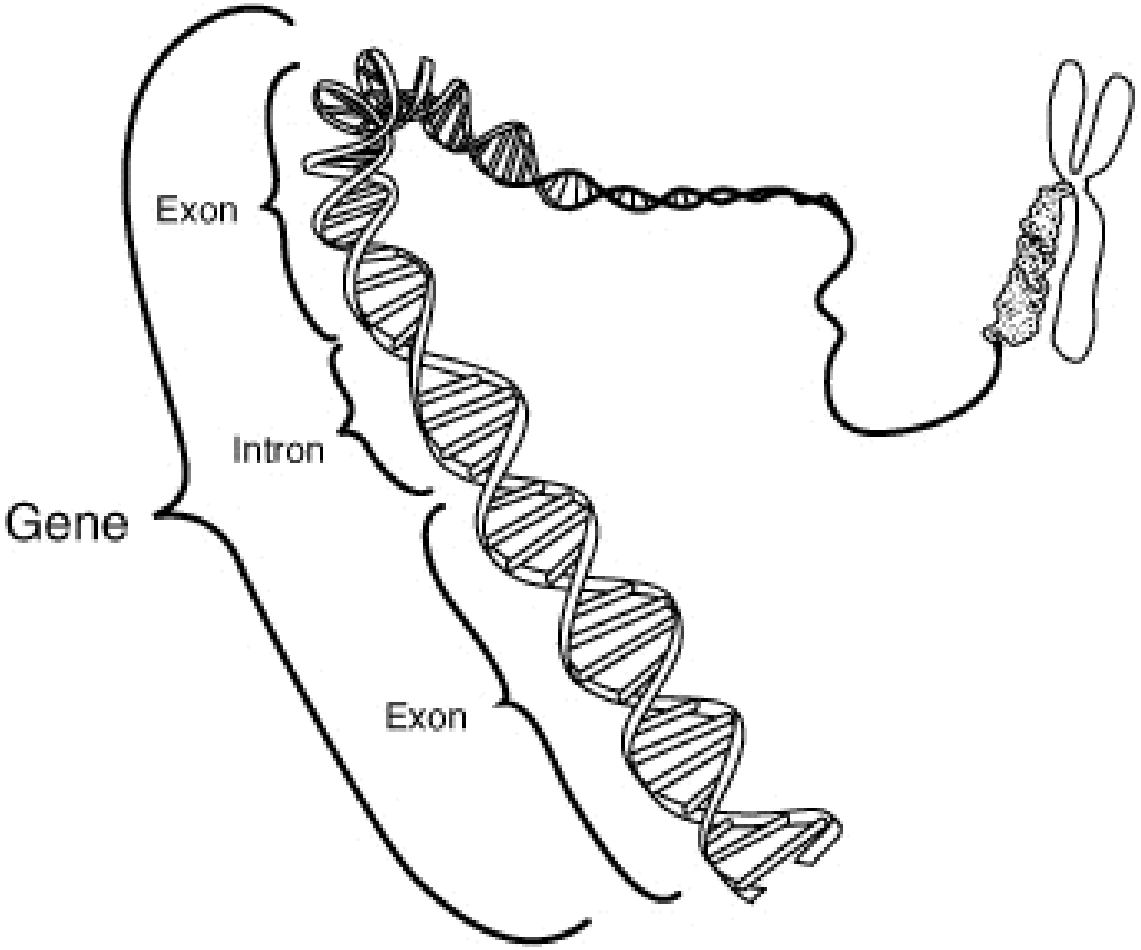
- Expression profiling represents a next step to sequencing a genome. The sequence tells us what the cell could possibly do, whereas the expression profile tells us what it is actually doing now.

- Gene expression profiling measures the activity of thousands of genes at once.

- The profiles can differentiate between cells that are actively dividing, or show how the cells react to a particular treatment.

- We will concentrate on tag-based techniques, e.g. serial analysis of gene expression (SAGE).

Gene Expression Profiling

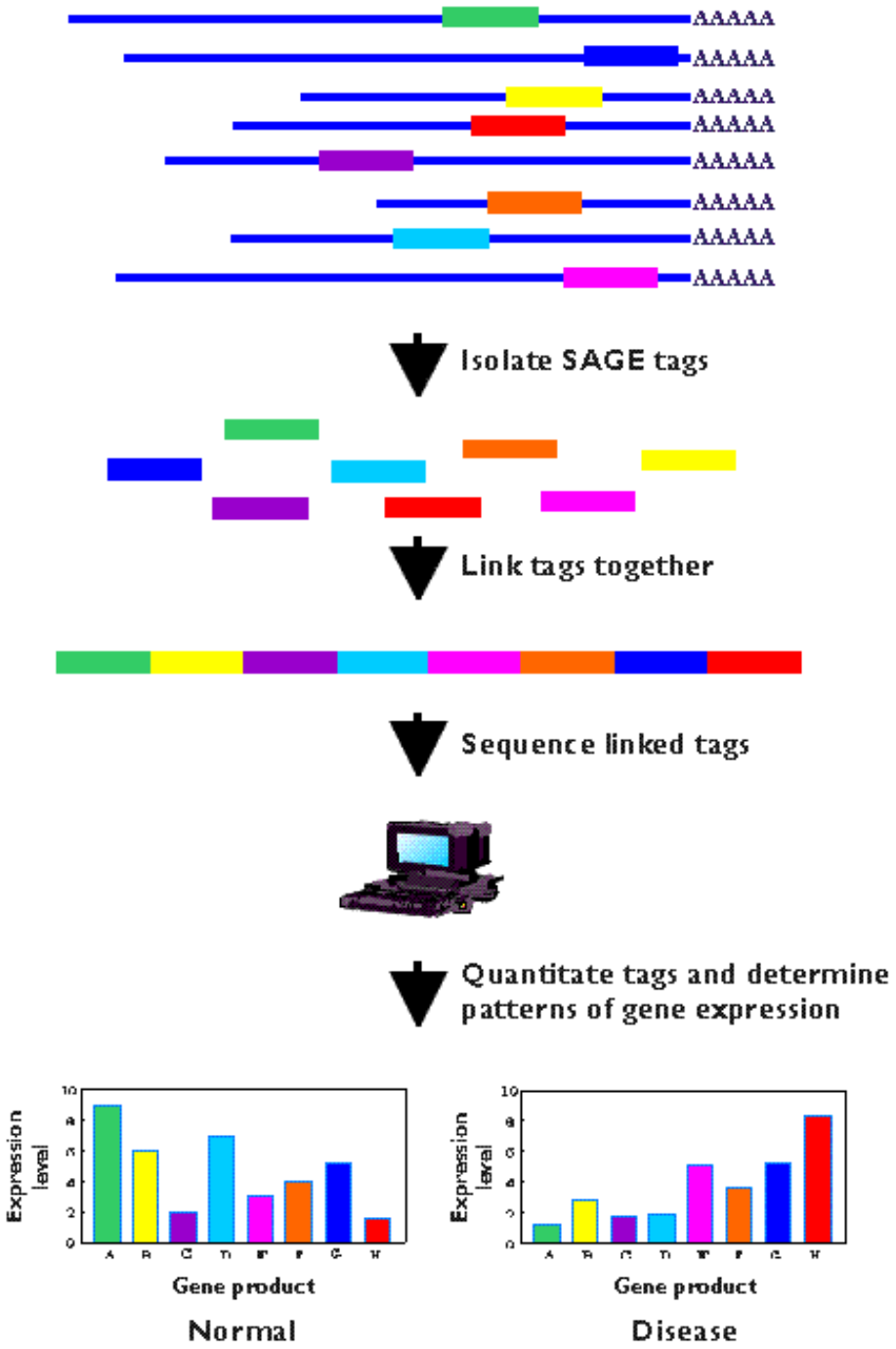


SAGE

Source: sagenet.org

- Serial analysis of gene expression (SAGE) is a technique used by molecular biologists to produce a snapshot of the messenger RNA population.

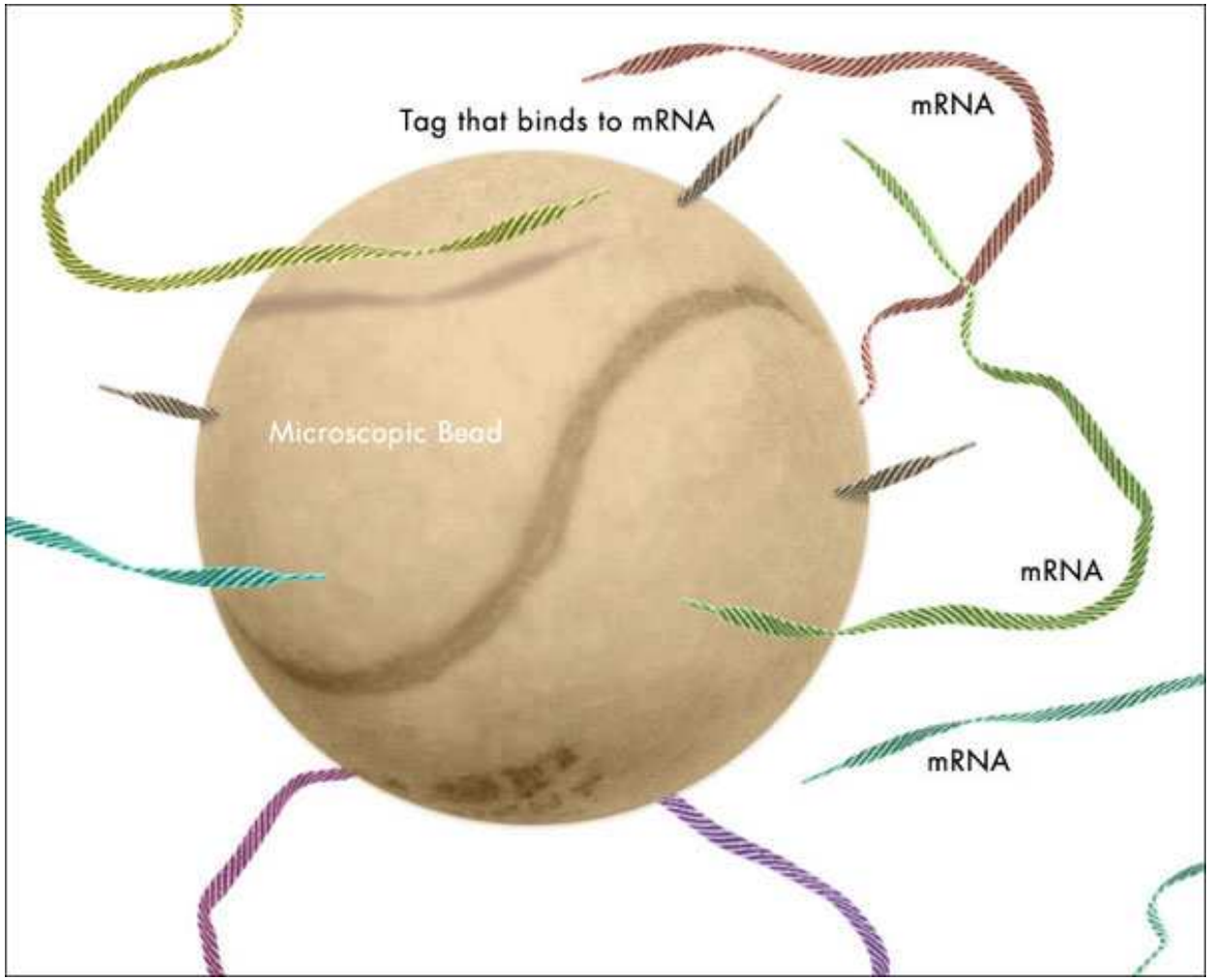
- Description of SAGE:
  1. A short sequence tag (10-14bp) contains enough information to uniquely identify a transcript (i.e. a strand of messenger RNA). The tag is obtained from a unique position within each transcript
  2. Sequence tags can be linked together to from long serial molecules that can be cloned and sequenced
  3. Quantitation of the number of times a particular tag is observed provides the expression level of the corresponding transcript.


AAAAA
AAAAA
AAAAA
AAAAA
AAAAA
AAAAA
AAAAA
AAAAA
Isolate SAGE tags
Link tags together
Sequence linked tags
Quantitate tags and determine patterns of gene expression
Expression level
Gene product
Normal
Expression level
Gene product
Disease

SAGE

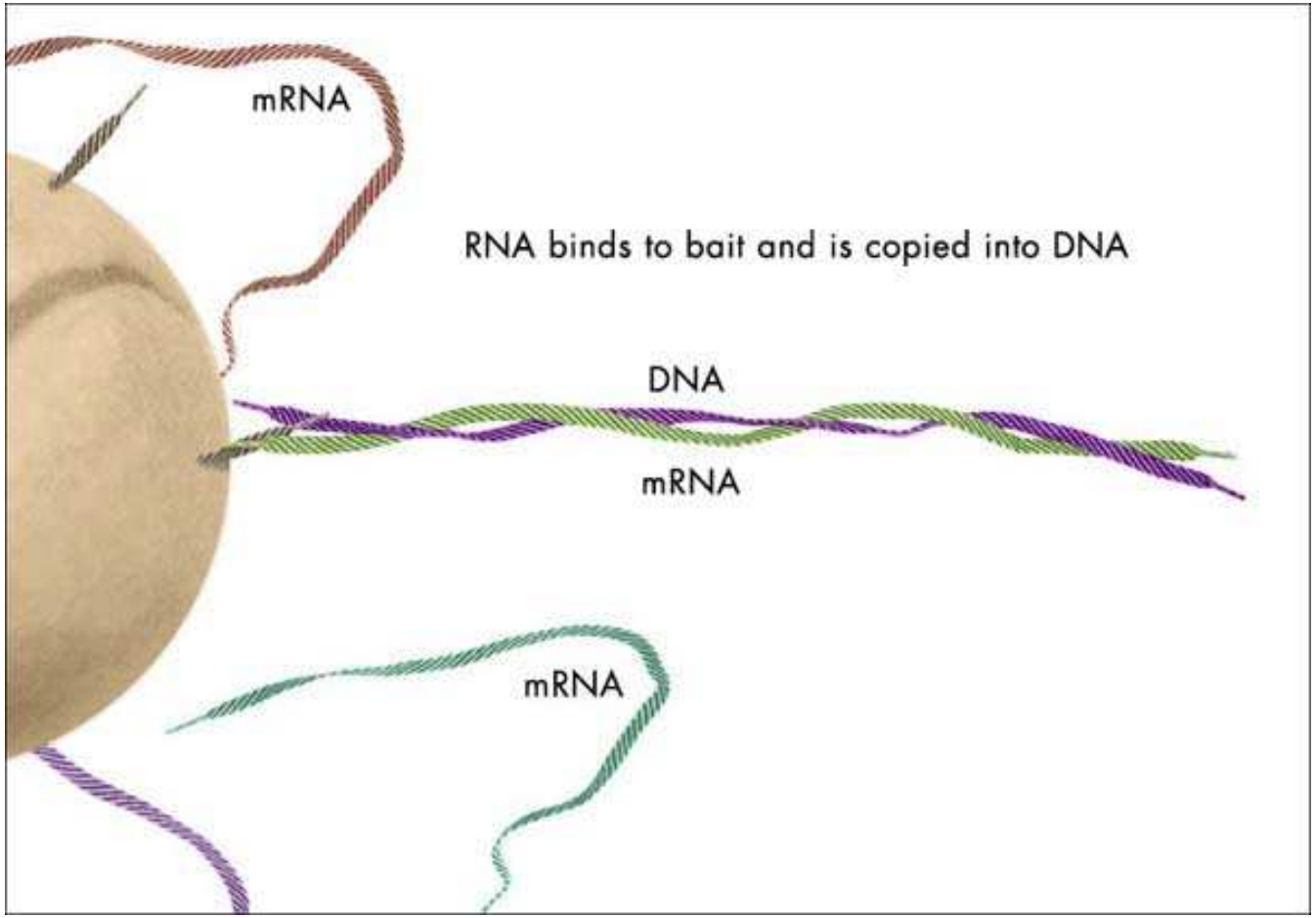
- SAGE experiments proceed as follows:
  - Isolate the mRNA of an input sample (e.g. a tumor).
  - Extract a small chunk of sequence from a defined position of each mRNA molecule.
  - Link these small pieces of sequence together to form a long chain (or concatemer).
  - Clone these chains into a vector which can be taken up by bacteria.
  - Sequence these chains using modern high-throughput DNA sequencers. This is where NGS comes in.
  - Process this data with a computer to count the small sequence tags.

SAGE



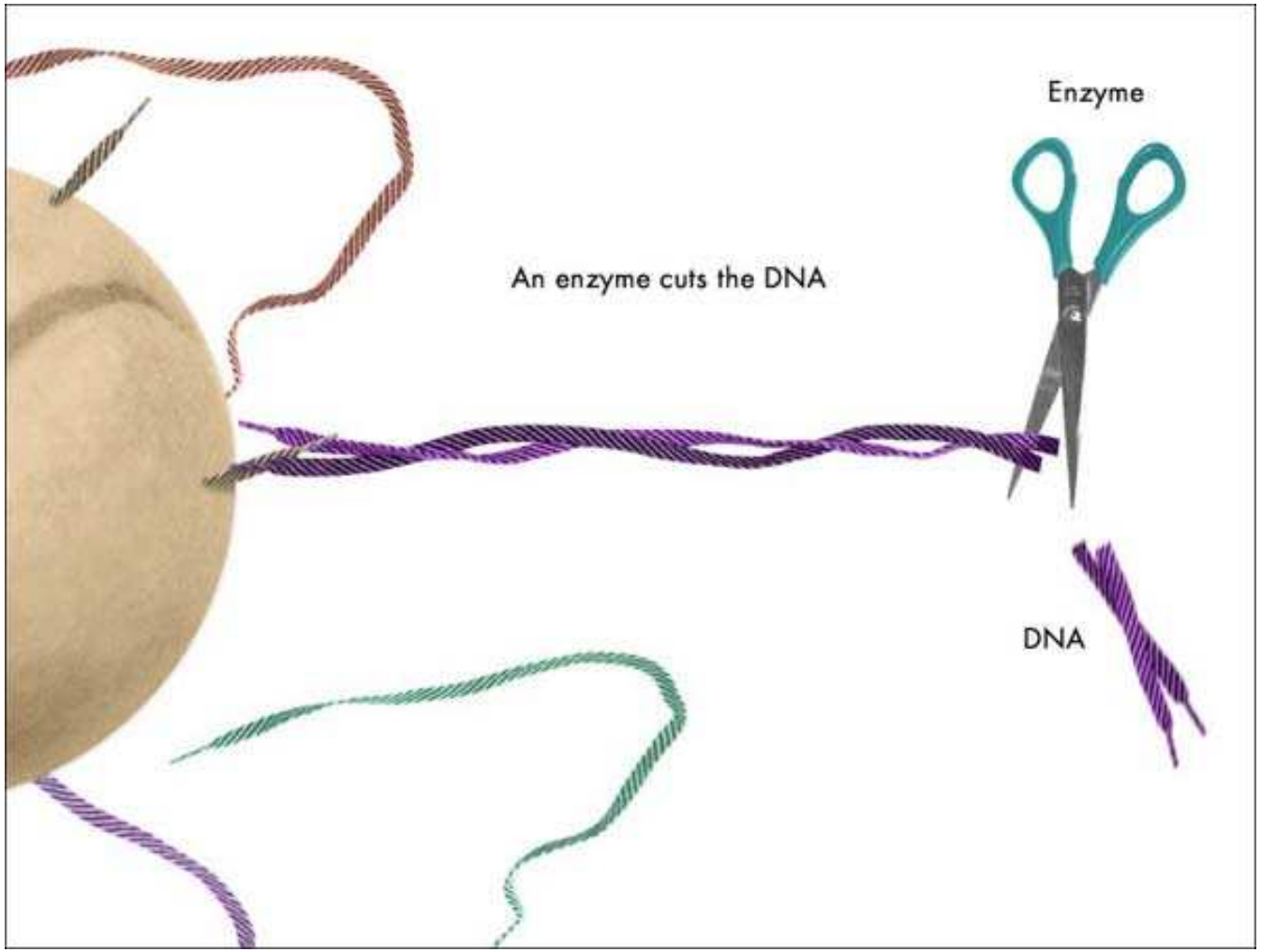
- Microscopic Bead and mRNA

SAGE



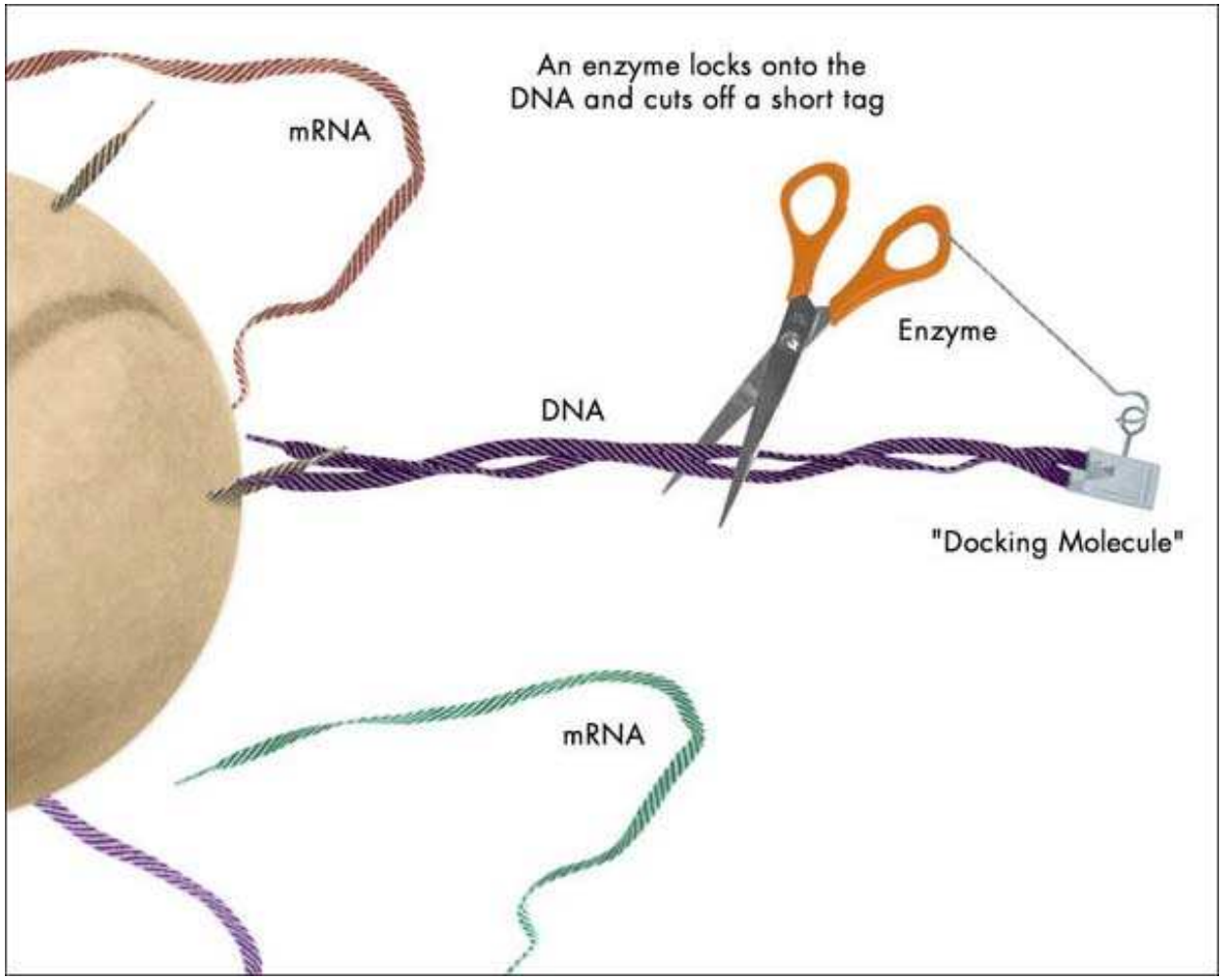
- RNA binds to bait and is copied into DNA

SAGE



- An enzyme cuts the DNA

SAGE



- An enzyme locks onto the DNA and cuts off a short tag

Restriction Enzymes

- There are hundreds of known restriction enzymes that cleave DNA at very specific sites.

- For example the enzyme BamHI recognizes the sequence GGATCC and cuts the DNA between the two G's.

  If just one base is changed in the sequence (say GGTTCC) then the enzyme will not cut the DNA.

- In the paper, three different restriction enzymes are used: Mbol, Nlalll, and Tail.

Conceptual Design

- The paper uses two different approaches to evaluate the performance of 454 sequencing.
- First approach: generated well-defined 3′ cDNA fragments by restriction enzyme treatment (as discussed above).
- The authors could predict the expected length of 3′ cDNA fragments, since they used a D. melanogaster strain with a fully sequenced genome.
- This strategy allowed them to evaluate whether fragment size affected 454 sequencing efficiency.

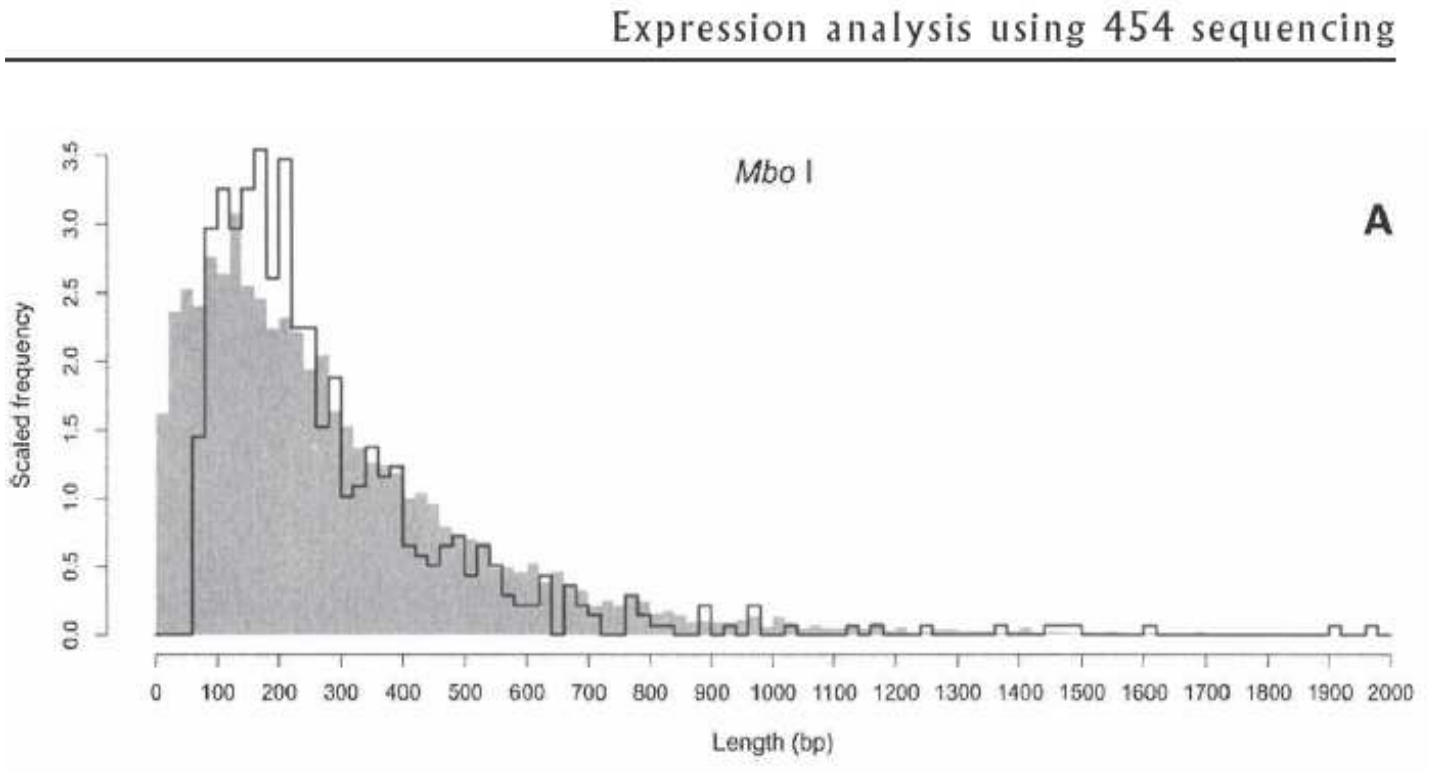
Conceptual Design

- Second Aproach: Randomly break 3′ cDNA fragments by high-pressure nitrogen (nebulization). This produces short DNA fragments for sequencing.
- The same mRNA was used for both approaches.
- This allowed comparison between the two different strategies and thus a measurement for the reliability of 454 sequencing-based expression profiling.

Biases in transcript representation

- As mentioned above, it was possible to predict the restriction fragment length of every known transcript.
- To compare this expected distribution to the observed one, every identified transcript was considered only once.
- Surprising Finding: For all enzymes tested, ESTs shorter than 80 bp or longer than 300 bp were under-represented

Biases in transcript representation



- The expected frequency distribution of 3′ cDNA fragment lengths is shown in grey. The black line indicates the frequency distribution obtained from 454 sequencing reads.

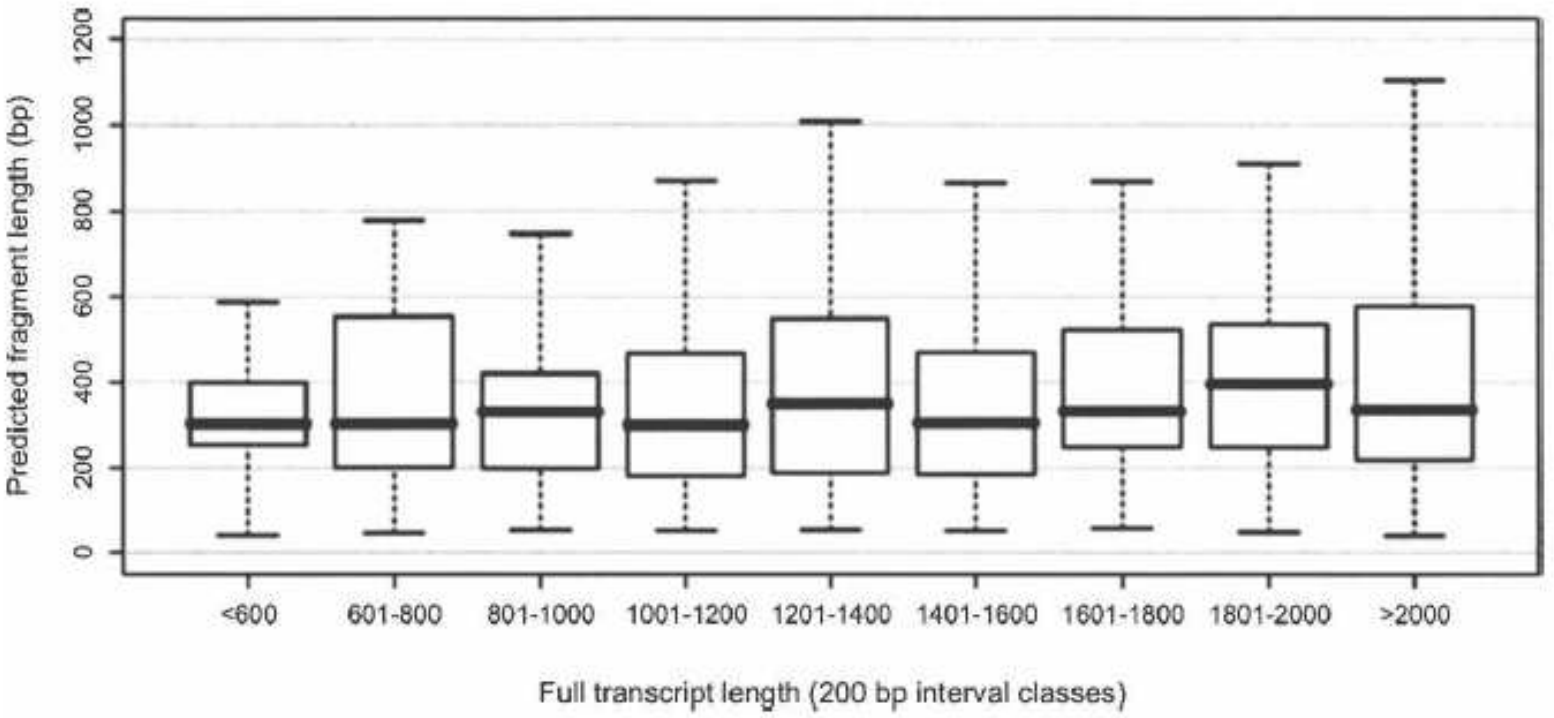
Nebulization success

- The size bias in the 454 sequencing could be overcome if every transcript had a similar distribution of fragment sizes.

- Idea: Randomly breaking cDNA fragments should overcome the size bias – affects all transcripts similarly.

- Shearing of DNA fragments by high-pressure nitrogen (nebulization) is frequently used to produce short DNA fragments for sequencing (Surzycki 2000).

Nebulization success



- Length distribution of 3′ cDNA fragments after nebulization among different size classes of full-length transcripts (as inferred from the available genome annotation).

Cross-method consistency

- Results of the nebulized cDNAs compared to cDNAs treated with restriction enzymes.

- When the expression levels of nebulized library were compared with the different digested libraries, the correlation coefficients ranged from 0.71 to 0.77.

Table 2. Consistency between libraries

Fragment length range considered	MboI	NlaIII	Tail
Full data set	**0.75**	**0.77**	**0.71**
<80 bp	0.89	0.94	0.93
80–300 bp	0.83	0.91	0.83
>300 bp	0.67	0.52	0.62

Correlation coefficients were calculated between the tag counts from the DIG library (as a whole and sorted) and counts from the NEB library.

Cross-method consistency

- The correlation coefficients were lower for the fragments longer than 300 bp – under-representation of long fragments.
- In fragments not suffering from an under-representation (80 – 300 bp), the correlation coefficients improved.
- The nebulized library showed a high correlation coefficient with each of the three different restriction libraries.
- These results indicate that the nebulization procedure is highly suitable to provide a reliable measurement of gene expression.

What we have so far

- Proof-of-principle study using D. melanogaster shows that the sequencing of randomly sheared 3′ cDNA provides a reasonable alternative to the previously suggested approaches.
- It would be also possible to sequence full-length cDNAs (Bainbridge et al. 2006; Emrich et al. 2007; Weber et al. 2007) rather than 3′ ends.
- However, the presented approach is more cost-effective.
- It requires only a single read per transcript and no adjustment for transcript length needs to be made.

Additional advantages

- The difference of the 454 sequencing technology to other massively parallel sequencing techniques – 454 produces longer reads. This is particularly important in the presence of sequence polymorphism.

- When evaluating the effect of short read lengths, about 20% of the 20-bp fragments had at least two perfect matches in the D. melanogaster genome.

- 50 and 100 bp fragments had substantially increased mapping accuracies – only 3% and 0.5% ambiguously mapped fragments, respectively.

- As expected, longer fragments result in a higher proportion of unambiguously mapped sequences

THE END