Rapid Transcriptome Characterization for a nonmodel organism using 454 pyrosequencing

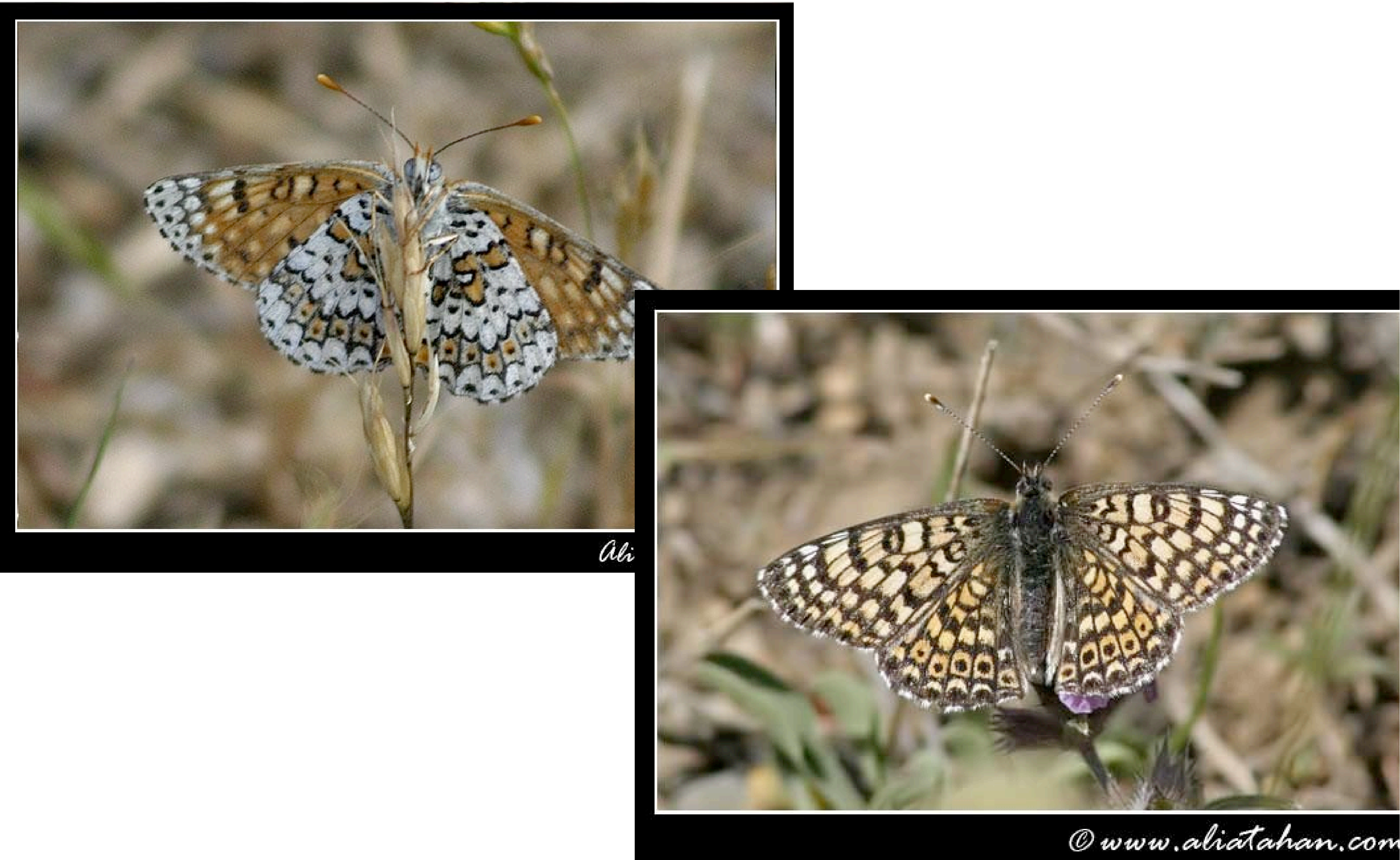
CRISTOBAL VERA, HRISTOPHER W. WHEAT, HOWARD W. FESCEMYER, MIKKO J. FRILANDER, DOUGLAS L. CRAWFORD, ILKKA HANSKI and JAMES H. MARDEN

Presented by Ilya Sutskever

The problem and the Paper

- Goal: Assemble the Transcriptomes/cDNA using NGS
  - Its **cheaper** than using Sanger
- Details:
  - Sequence cDNA with 454 and Sanger
  - Show that the 454 is useful for many tasks, and is no worse than Sanger (but cheaper).

The subject: Glanville Fritiliy butterfly

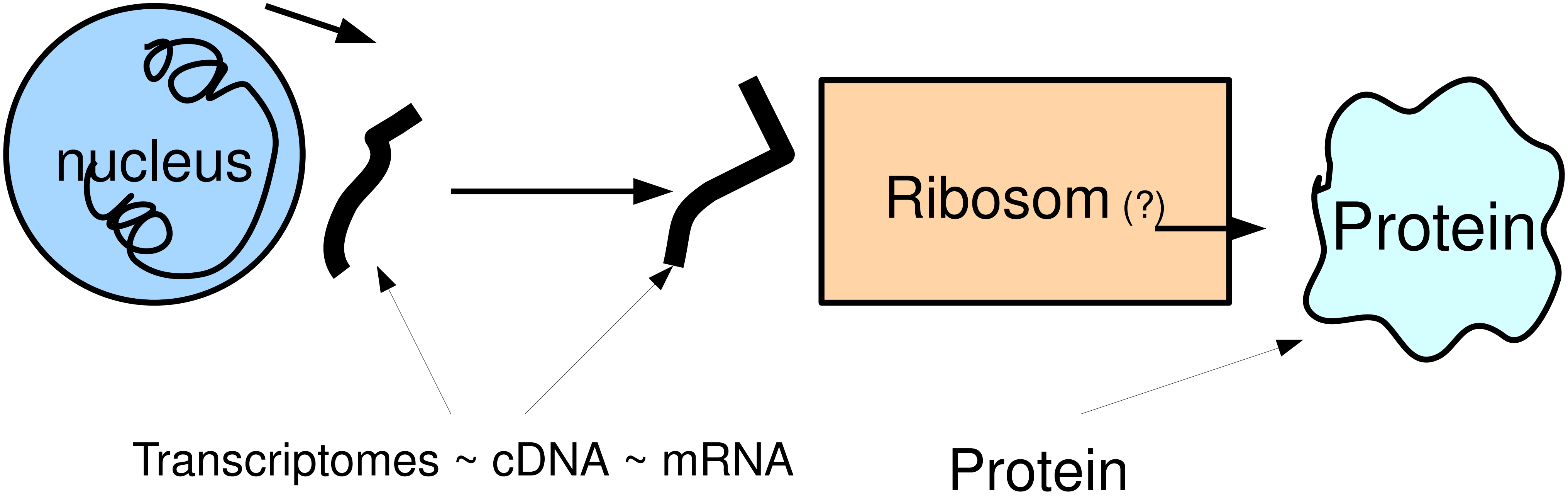


Recap: 454 and Sanger

- 454:
  - 4.5 hours
  - $2K
  - Read length: 110 bp
  - 300,000 reads
  - ~ 30 Mbase
- Sanger: expensive:
  - Read length: 500bp

Transcriptomes and cDNA

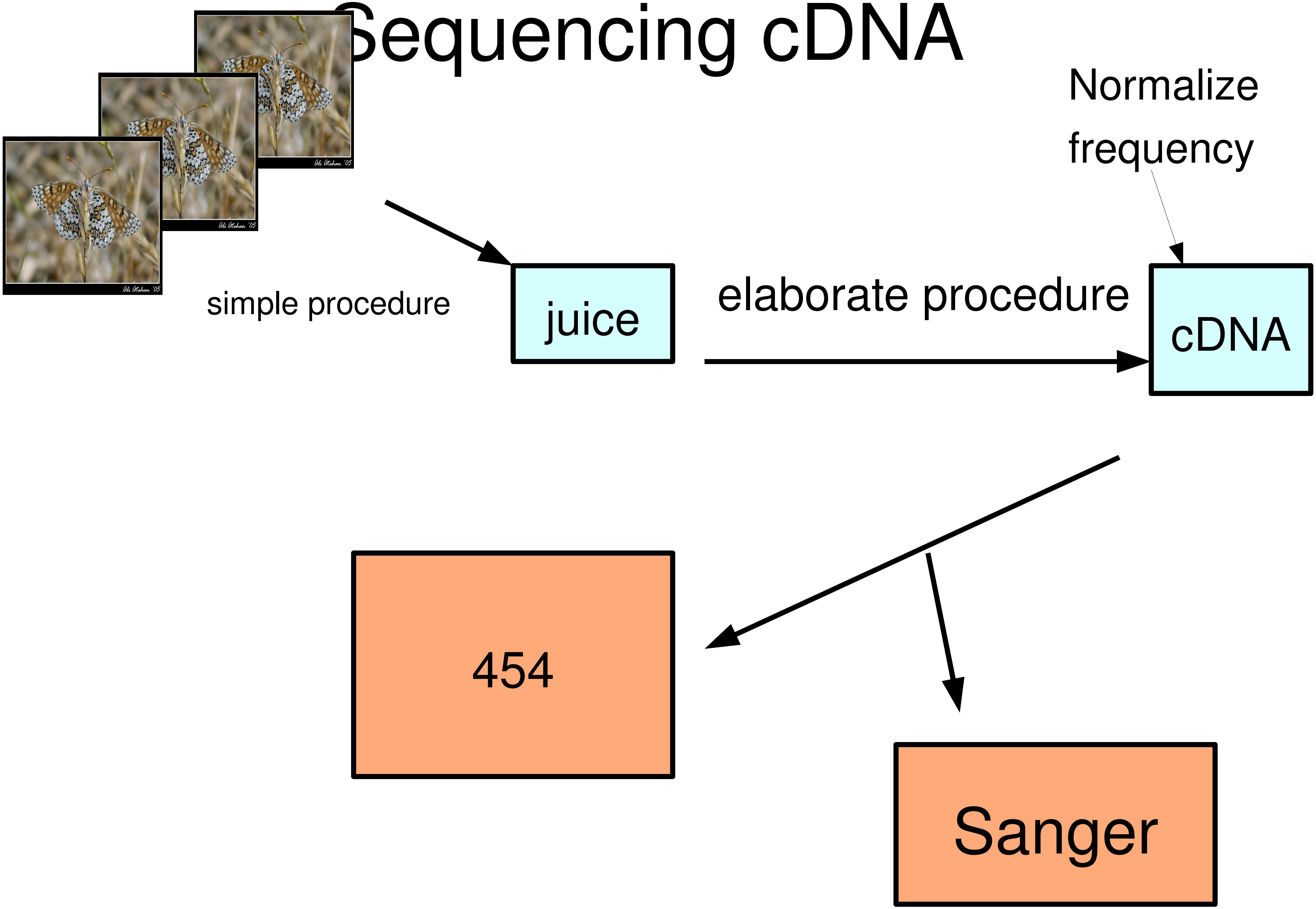
- (I think that) these are the DNA sequences that are currently used to generate proteins.
- They correspond to the expressed proteins.



Comparison to previous work

- 454 was used before for transcriptome sequencing
- But ...
  - Either Sanger was also used or a reference genome was known
  - Or lower coverage was used, so assembly was impossible
-

Sequencing cDNA



Ali Atahan '05

Details of the process

- Get RNA from larvae, pupae, and from adults.
  - From a diverse population
  - The butterfly will have different transcriptomes in different stages of its life
- RNA -> cDNA (magic)

Algorithm

- SEQMAN PRO 7.1
  - Use it to get rid of low quality data
  - Use it to assemble the reads from Sanger and from the 454 – get contigs.
  - That's it.

What to do with the data?

- Take a database of proteins, Uniprot 9.2
- Align the contigs to the proteins, to find which proteins are expressed in the butterfly
- More alignments to proteins of :
  - *Bombyx mori*
  - *Drosophila melanogaster*
  - *M. cinxia*
  - Butterflybase

Microarrays

- Some good contigs (ones that matched good proteins, I think) were used as probes for microarrays
- 200K microarray probes were generated
- Microarrays tell us what genes are expressed

Results of sequencing

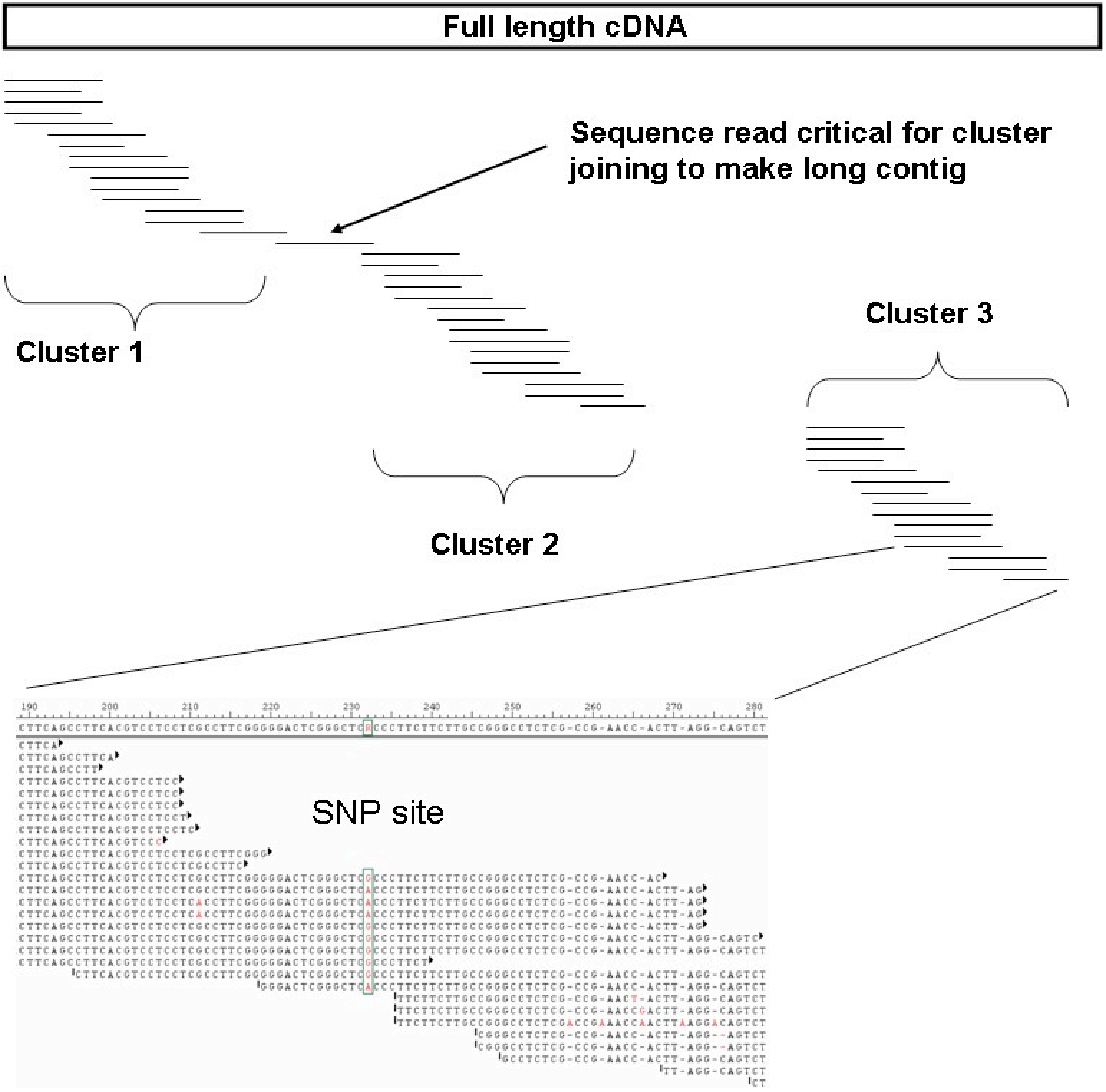
- 50K contigs, mean length 200 bp (it seems short to me)
- They tried to look for exact matches between contigs. But most of these matches matched to different proteins (except 2%)
- So these must be motifs in different proteins

Sanger vs 454

- 92% of Sanger reads had strong alignments to 454 contigs
- Contigs had very few gaps when aligned to Sanger

Coverage is important for assembly

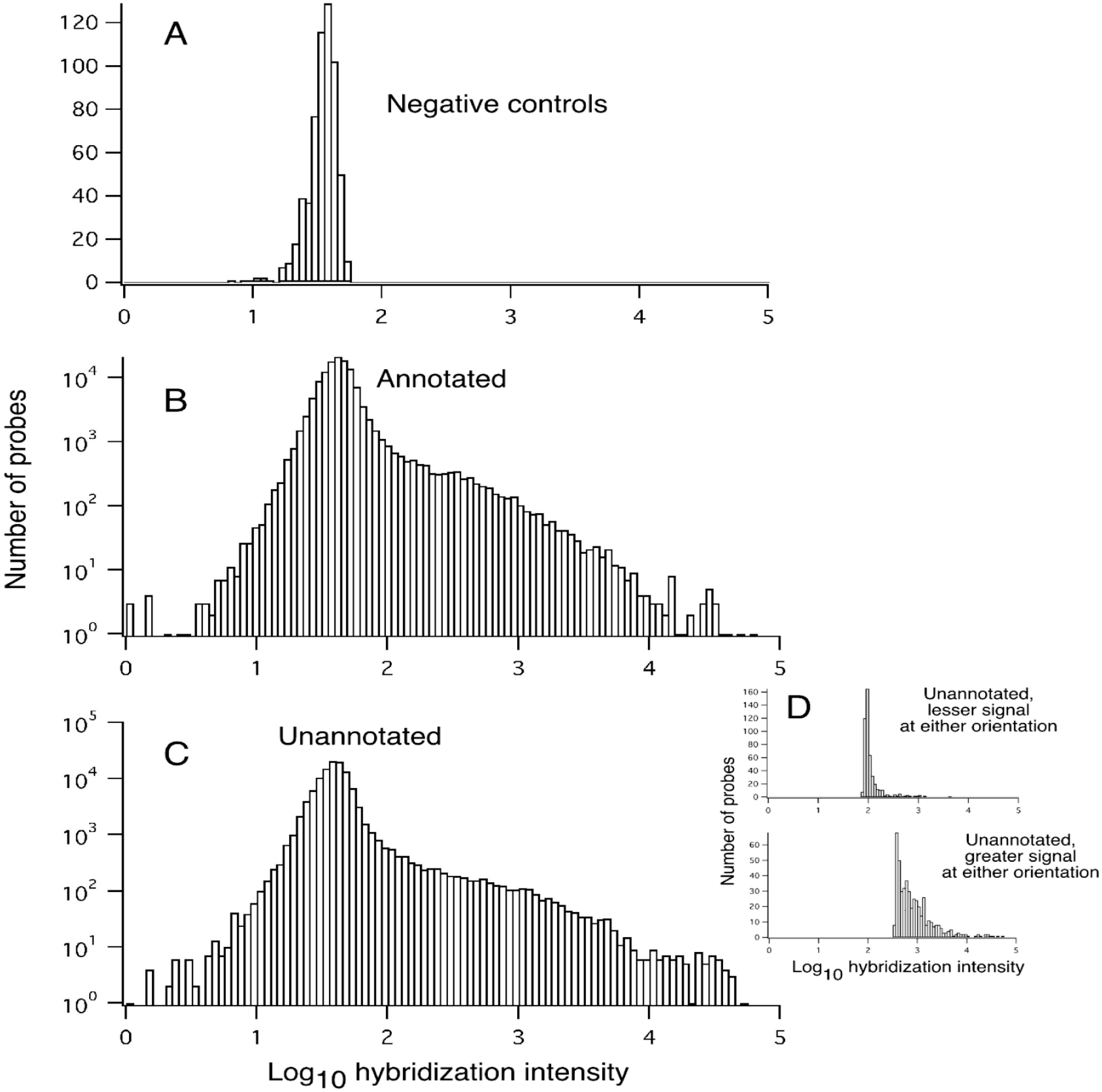
- They have evidence for that.



```
CTTCAGCCTTCACGTCCTCCTCGCCTTCGGGGGACTCGGGCTCRCCCTTCTTCTTGCCGGGCCTCTCG-CCG-AACC-ACTT-AGG-CAGTCT
CTTCA
CTTCAGCCTTCA
CTTCAGCCTT
CTTCAGCCTTCACGTCCTCC
CTTCAGCCTTCACGTCCTCC
CTTCAGCCTTCACGTCCTCC
CTTCAGCCTTCACGTCCTCCT
CTTCAGCCTTCACGTCCTCCTC
CTTCAGCCTTCACGTCCC
CTTCAGCCTTCACGTCCTCCTCGCCTTCGGG
CTTCAGCCTTCACGTCCTCCTCGCCTTC
CTTCAGCCTTCACGTCCTCCTCGCCTTCGGGGGACTCGGGCTCGCCCTTCTTCTTGCCGGGCCTCTCG-CCG-AACC-AC
CTTCAGCCTTCACGTCCTCCTCGCCTTCGGGGGACTCGGGCTCACCCTTCTTCTTGCCGGGCCTCTCG-CCG-AACC-ACTT-AG
CTTCAGCCTTCACGTCCTCCTCACCTTCGGGGGACTCGGGCTCACCCTTCTTCTTGCCGGGCCTCTCG-CCG-AACC-ACTT-AG
CTTCAGCCTTCACGTCCTCCTCACCTTCGGGGGACTCGGGCTCACCCTTCTTCTTGCCGGGCCTCTCG-CCG-AACC-ACTT-AG
CTTCAGCCTTCACGTCCTCCTCGCCTTCGGGGGACTCGGGCTCGCCCTTCTTCTTGCCGGGCCTCTCG-CCG-AACC-ACTT-AG
CTTCAGCCTTCACGTCCTCCTCGCCTTCGGGGGACTCGGGCTCGCCCTTCTTCTTGCCGGGCCTCTCG-CCG-AACC-ACTT-AGG-CAGTC
CTTCAGCCTTCACGTCCTCCTCGCCTTCGGGGGACTCGGGCTCGCCCTTCTTCTTGCCGGGCCTCTCG-CCG-AACC-ACTT-AGG-CAGTCT
CTTCAGCCTTCACGTCCTCCTCGCCTTCGGGGGACTCGGGCTCGCCCTTCT
       CTTCACGTCCTCCTCGCCTTCGGGGGACTCGGGCTCGCCCTTCTTCTTGCCGGGCCTCTCG-CCG-AACC-ACTT-AGG-CAGTCT
                              GGGACTCGGGCTCACCCTTCTTCTTGCCGGGCCTCTCG-CCG-AACC-ACTT-AGG-CAGTCT
                                               TTCTTCTTGCCGGGCCTCTCG-CCG-AACT-ACTT-AGG-CAGTCT
                                               TTCTTCTTGCCGGGCCTCTCG-CCG-AACCGACTT-AGG-CAGTCT
                                               TTCTTCTTGCCGGGCCTCTCGACCGAAACCAACTTAAGGACAGTCT
                                                         CGGGCCTCTCG-CCG-AACC-ACTT-AGG--AGTCT
                                                         CGGGCCTCTCG-CCG-AACC-ACTT-AGG--AGTCT
                                                            GCCTCTCG-CCG-AACC-ACTT-AGG-CAGTCT
                                                                                TT-AGG-CAGTCT
                                                                                           CT
```

Transcriptome coverage Breadth

- 20% of the contigs were well aligned to proteins in the different databases
- 9000 unique proteins were detected this way
  - with 73% amino acid identity
- If we microarray some of the unmatched reads, the responsiveness of the microarray is the same for annotated and unannotated (matched) contigs. So more proteins were found.


A
Negative controls
120
100
80
60
40
20
0
0
1
2
3
4
5
B
Annotated
Number of probes
10^4
10^3
10^2
10^1
10^0
0
1
2
3
4
5
C
Unannotated
10^5
10^4
10^3
10^2
10^1
10^0
0
1
2
3
4
5
Log$_{10}$ hybridization intensity
D
Unannotated,
lesser signal
at either orientation
160
140
120
100
80
60
40
20
0
0
1
2
3
4
5
Unannotated,
greater signal
at either orientation
60
50
40
30
20
10
0
0
1
2
3
4
5
Number of probes
Log$_{10}$ hybridization intensity

Functional annotation

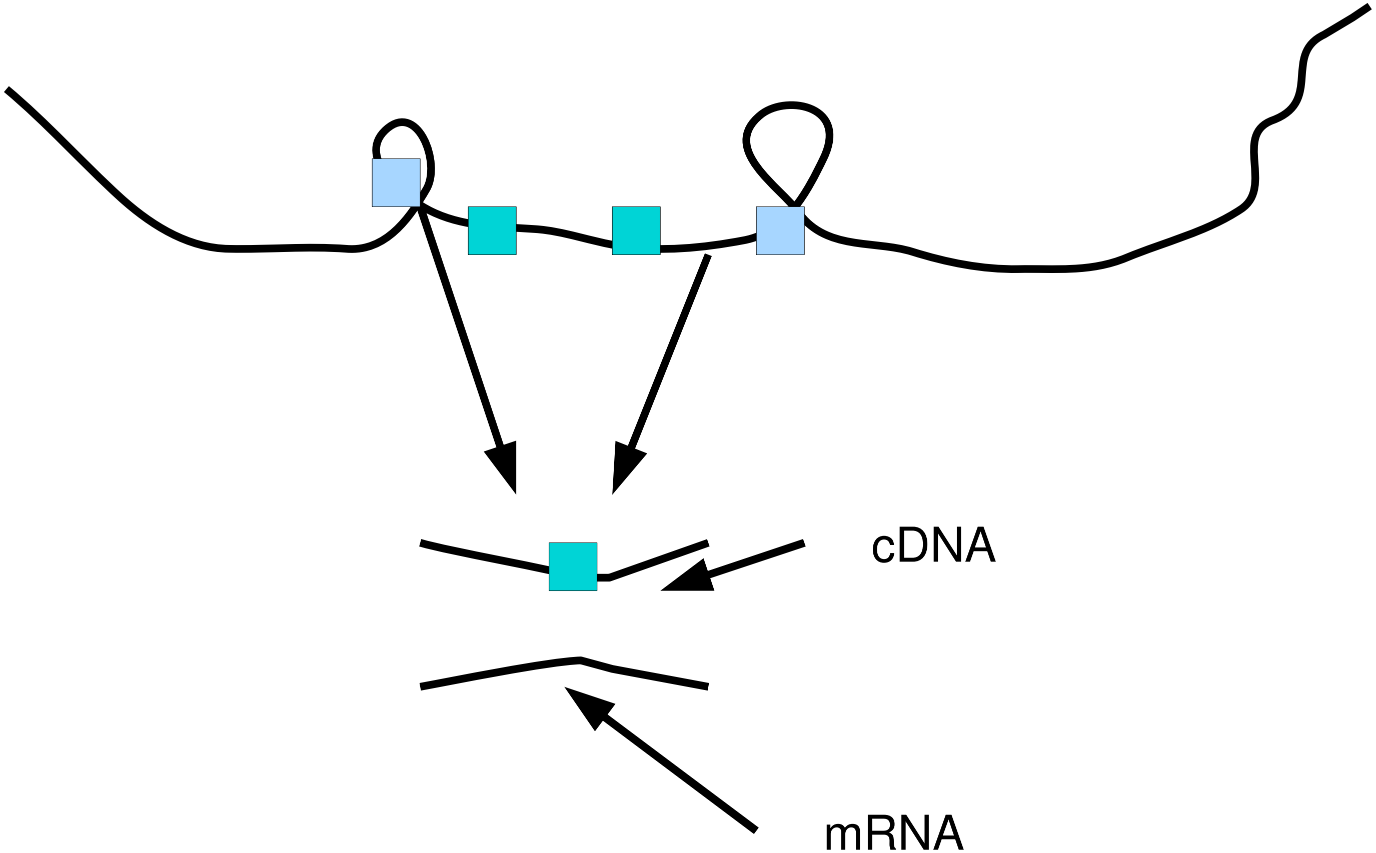
- Not too sure...
- The reads/contigs were matched to known proteins with known function
- This way, the function of the reads was guessed

SNP discovery

- Take the contigs, and discover SNPs
- 6.7 SNPs per 1000 base pairs
- 751 SNPs at 6X covered sites, in 355 contigs

Alternative splicing

- It is when the dna is spliced before turning to cDNA and mRNA



Alternative splicing effects on assembly

- Characterize 2 such genes using PCR, cloning method, amplification of cDNA ends
- The genes have deep coverage
- Somehow, it made things more difficult

Detection of intracellular parasite

- Many reads had alignment to sequences of non-insects
- That's pretty much it!