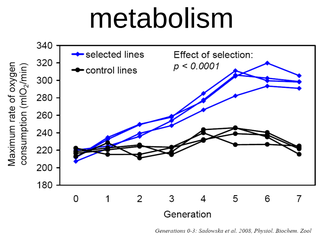
Investigating molecular basis of response to selection in the bank vole with next generation sequencing

M. Stuglik[1], W. Babik[1], W. Qi[2], M. Kuenzli[2], K. Gac[1], P. Koteja[1], J. Radwan[1]

[1]Institute of Environmental Sciences, Jagiellonian University, ul. Gronostajowa 7, 30-387 Krakow, Poland
[2]Functional Genomics Center Zurich, Winterthurerstr. 190, 8057 Zurich, Switzerland.

Aims

- characterize heart transcriptome of the bank vole at the sequence level
- catalogue genes expressed in heart
- check for SNPs in coding regions differentiating regimes in a selection experiment



Methods

Total RNA from multiple individuals per line (4 lines selected for high metabolism (Selected) and 4 control lines (Control)) was extracted from heart using RNAeasy kit (Qiagen).
cDNA produced using MINT (Evrogen) protocol modified to disrupt poly A/T tails to facilitate 454 sequencing.
Double stranded cDNA normalized using Trimmer kit (Evrogen).
Equal amounts of normalized cDNA from each line combined into Selected and Control pools which were sequenced in two halves of a 454 Titanium run.
Reads were assembled into contigs using CAP3 and SNPs were detected in .ace files using GigaBayes
Both contigs and singletons were used for searching SwissProt protein database and mouse RefSeq mRNA database.
12 SNPs differing significantly between selection regimes were validated in genomic DNA (gDNA) of five individuals per line (40 individuals in total) using the SNaPshot method.

Table 1. Basic statistics of the 454 run

Raw bases	351.6 Mb
Raw reads	1.1 mln
Mean raw read length ± SD	316.8 ± 127.2 bp
Median raw read length	348 bp
Trimmed bases	311.0 Mb
Mean trimmed read length ± SD	280.2 ± 130.9 bp
Median trimmed read length	310 bp

Table 2. Characteristics of contigs

N	63,348
Min length	96 bp
Max length	13,292 bp
Mean length ± SD	481.0± 294.4 bp
Median length	417 bp
Max coverage	20,753
Mean coverage ± SD	10.7 ± 107.4
Median coverage	3

Fig. 2. The frequency distribution of CAP3 contigs length.

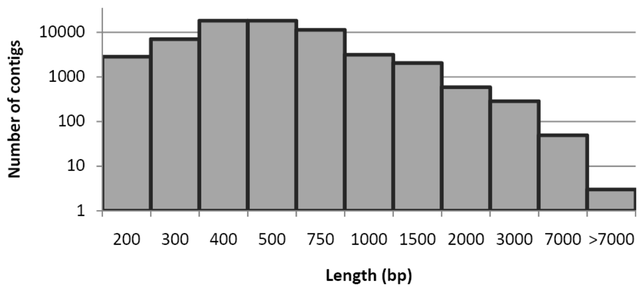


Fig. 1. The distribution of maximum transcript length (based on the mouse RefSeq mRNA) for genes detected in tha bank vole compared to the distribution for all genes in the mouse RefSeq. **Note over-representation of long transcripts in the bank vole data**

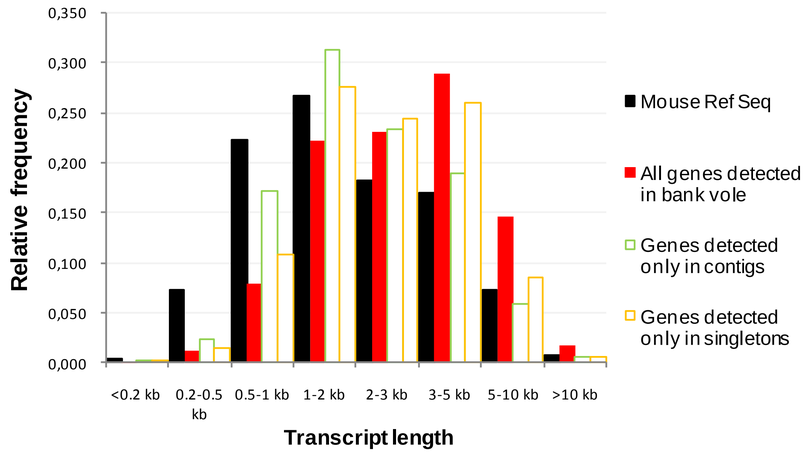


Table 3. Sequences matching known genes (BLAST threshold E-value 10^{-5})

	SwissProt	Mouse RefSeq mRNA
Percent of contigs with hits	29.4	43.0
Percent of singletons with hits	12.8	28.0
N of unique genes detected	11,249	14,607
N of unique genes with matches only in contigs	1,601	1,145
N of unique genes with matches only in singletons	3,349	4,668

Fig. 3. Frequencies of genes in six Gene Ontology categories: all genes detected, and genes containing SNPs differentiating selection regimes. Asterisks indicate significant overrepresentation of genes containing SNPs in comparison to all genes in respective categories (chi sqare test).

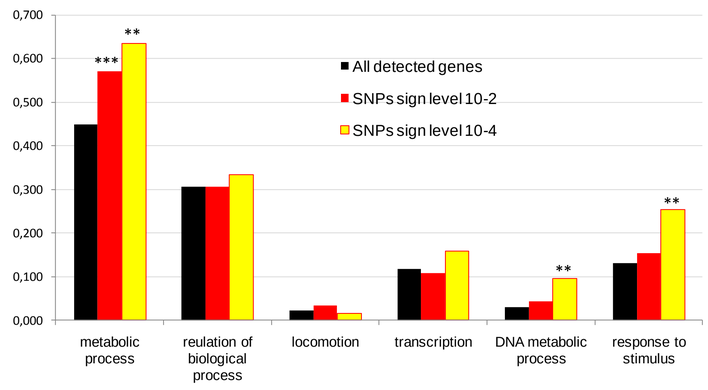
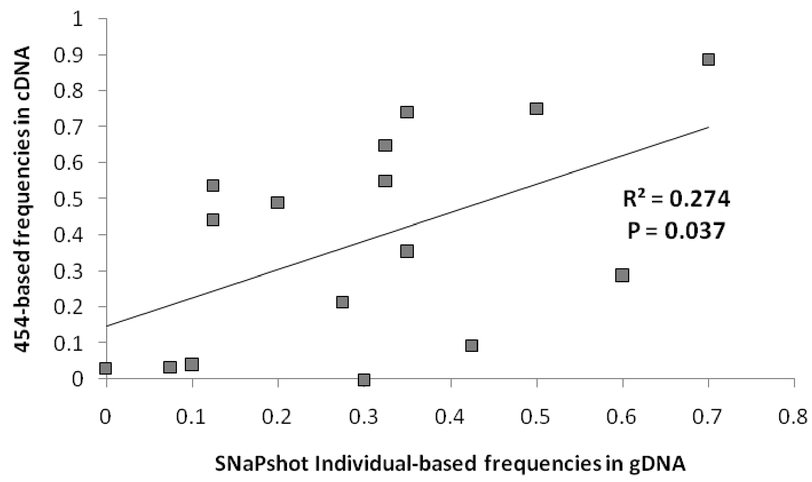


Fig. 4. Correlation between the frequency of a given SNP in 454 pools (E and C) and individual-based frequencies in pools obtained from SNaPshot gDNA genotyping of 20 individuals per pool (five voles = 10 alleles per line).



Results

Transcriptome characterization

Titanium chemistry enabled obtaining longer reads (Table 1), longer contigs (Table 2, Fig. 2) than in most previous studies using FLX chemistry and detection of over 12,000 genes (Table 3)
Our representation of the transcriptome is comprehensive as evidenced by the complete or almost complete representation of several macromolecular complexes (in parentheses number of known genes): 26S proteasome (22), Chaperonin (8), Nuclear Pore (28), Respiratory chain complex I (38), Ribosoome (79), spliceosome (134 out of 141)
and metabolic pathways: glycolysis (10), gluconeogenesis (10), pentose phosphate (6), citric acid (14).
Our PCR-based laboratory procedures apparently did not select against long transcripts (Fig. 1).

SNP differences between selective regimes

22,471 SNPs with posterior P > 0.99.
136 SNPs in 81 contigs had significantly different frequencies between selection regimes at P = 10^{-4} and 1388 SNPs in 736 at P = 10^{-2}.
Genes involved in metabolic processes and response to stimulus appear to be overrepresented (Fig. 3).
Twelve SNPs were validated experimentally, nine were confirmed, and one found significantly different between selection regimes on the basis of per-lineage mean frequencies, there was however significant correlation between frequencies of SNPs in selection regimes inferred from 454 data and from individual based genotyping (Fig. 4).

Conclusions

- A single 454 Titanium run enabled comprehensive characterisation of heart transcriptome (>12000 genes detected)
- >22000 candidate SNPs detected; 75% (9/12) experimentally validated
- Candidate SNPs responsible for divergence between selection regimes identified