Sympatric speciation of spiny mice, *Acomys*, unfolded transcriptomically at "Evolution Canyon", Israel

Kexin Li, Huihua Wang, Zhenyuan Cai, Liuyang Wang, Qinqin Xu, Matěj Lövy, Xiaoying Song, Zhenlong Wang, Eviatar Nevo

This PDF file includes:

Supplementary Tables S1

Supplementary Figures S1-S7

Supplementary Appendix Suggested Readings

SI Appendix Materials and Methods

Ethics statement

All the experiments on *Acomys* were conducted following the rules and guidelines of the University of Haifa and approved by the Ethics Committee.

Sample collection

Seven stations were defined on the two slopes of EC (**Fig. 1**). Station #2 on the south-facing "African" savannoid slope (AS) of EC I represents the environment with high solar radiation, temperature, and drought. By contrast, station #6, on the north-facing "European" forested slope, separated on average by 250 meters, represents the microclimate of lower solar radiation and cool temperature, and higher humidity (1, 2). Six and four animals (**Table S1**) were collected from station #2 and #6 at Evolution Canyon I (EC I), lower Nahal Oren, Mount Carmel, Israel respectively, in September 2013. They were sacrificed by injection with Ketaset CIII (Fort Dodge, USA) at 5 mg/kg of body weight after left in the lab for about 5 hours. Whole brain tissues were harvested and immersed immediately into liquid nitrogen.

Table S1 Morphological information for all samples of *Acomys cahirinus* from EC I

Sample	Body weight	Tail length	Body length	Sex
AS1-20				
AS1-21				
AS2-1	38.8		10.8	
AS2-2	36.3	9.7	10	♂
AS2-4	47.6	8	10.8	♀
AS2-8	27.3	9.1	10.1	
ES6-5	46.9	10.4	11.2	♀
ES6-7	41.4	9.8	11	♀
ES6-8	30.4	9.6	10.1	
ES6-13	30.4	9.4	9.9	

cDNA library construction and sequencing

Total RNA from the whole brain was extracted from each individual using an RNA kit (Qiagen, Germany) according to the manual. The quantity and quality of the RNA were assessed by Qubit® 3.0 fluorometer (ThermoFisher Scientific, USA), 1% agarose gel electrophoresis and Agilent 2100 (Agilent Technologies, USA), and only the samples showing RNA integrity number (RIN) > 7 were used for downstream experiments. Pair end libraries with about 300 bp insert size were constructed with a unique barcode for each sample using Illumina Tru-seq RNA kit Sample Preparation Kit v2 (Illumina, San Diego, CA) according to the manufacturer's protocols. Briefly, 4 ug total RNA was purified with Sera-mag oligo-dT attached magnetic beeds (Illumina, San Diego, CA). The isolated mRNA was fragmented with divalent cations fragmentation buffer incubated at 94 ℃ for 5 min. The first strand cDNA was synthesized based on mRNA and random primers. The second strand cDNA was synthesized using DNA polymerase I and RNase H and dNTPs followed by AMPure XP beads purification. cDNA fragments were subjected to end repair, 3’ adenylation, and ligation of the adaptor. The final library was generated following PCR amplification and purification. After quantification by Qubit® 3.0 fluorometer (ThermoFisher Scientific, USA) and qPCR (LightCycler® 480 Instrument II, Greece), cluster generation was conducted on cBot (Illumina, USA), and later 100~150bp pair end sequencing was conducted on HiSeq 2000 **(**Illumina, San Diego, CA).

Transcriptome assembly

All reads from the 10 samples were sorted by the barcode. Adaptor sequences were clipped and low-quality reads, including putative PCR duplicates, and the reads that their average base quality was < 20 and those with >5% unidentified nucleotides (N) were filtered out by Fast-Tool kit (3). Base quality was checked and visualized by FASTQC (4). The left and right clean reads from the 10 individuals were used as the left and right input into Trinity (5), respectively, with default parameters but min_kmer_cov 2. The transcripts generated from Trinity were clustered by CAP3 and later filtered by CD-HIT (6) with default parameters.

Transcriptome annotations

All of the unigenes were filtered and only the longest transcript from each unigene was kept. Sequence homology searches were carried out using BLAST programs against sequences in NCBI non-redundant protein sequences (Nr) (E-value=10^{-5}), Blast against NCBI nucleotide sequences (Nt) (E-value=10^{-5}), Swiss-prot (E-value=10^{-5}), "KEGG (Kyoto Encyclopedia of Genes and Genomes)" (E-value=10^{-3}), and COG (Cluster of Orthologous Groups of proteins) (E-value<10^{-3}). The unmapped unigenes were predicted by ESTscan (7). The length distribution of unigenes was estimated by homemade script. Blast2GO (8) was used for gene ontology (GO) analysis with an E-value of 1e-6.

Variant calling

Single nucleotide polymorphism (SNP) was called by GATK2 (9). The clean reads from each individual were mapped to the transcript consensus using BWA (10). The generated SAM files were reordered after index building with Picard tools (http://picard.sourceforge.net). SAM files were transformed to binary BAM files using SAM tools (11), which was followed by BAM files sorting and head adding. Duplicates were masked and the generated files were indexed again. In order to minimize the mapping error, local realignment around INDELs were conducted again using Picard tools. Base Recalibration was conducted in case of systematic error modes. Raw SNP datasets were filtered with the following parameters: cluster Window Size: 10; MQ0 >= 4 and (MQ0/(1.0*DP)) > 0.1; QUAL < 10; QUAL < 30.0 or QD < 5.0 or HRun > 5), and only SNPs with distance > 5bp were retained for downstream analysis. Hardy-Weinberg equilibrium (HWE) was tested with VCFtools (http://vcftools.sourceforge.net/) and those deviating from HWE (P<0.05) were removed from downstream analyses.

The loci with more than two alleles or two missing genotypes were removed by perl script, and 126,074 SNPs were retained. Finally, SNPs that could not pass the following two criteria were excluded: (1) SNPs with minor allele frequency (MAF) > 0.01; (2) maximum per-SNP missing rate < 0.1. After this step, there were 73,418 SNPs in the genetic diversity analysis dataset.

Standard population genetic statistics, including Watterson's θ, and pairwise nucleotide diversity π were calculated for each population by the Bio::PopGen::Statistics package in BioPerl (v1.6.1) (12).

Population analyses

EIG4.2 software was used to conduct PCA on the SNP dataset (13). Genetic structure was inferred using ADMIXTURE 1.23 (14), which implements a block-relaxation algorithm. Default parameters were used in Admixture analysis. Matrix pairwise F_{ST} value was estimated for all loci between populations using the Genepop 4.2.2 software (15), then rescaling F_{ST} as F_{ST} /(1- F_{ST}), and the neighbor-joining tree for populations were constructed with R package *ape* based on matrix pairwise rescaling F_{ST} values (16). We also constructed the neighbor-joining tree for individuals using SplitsTree software (17). The kinship between individuals were calculated by KING software (18). The heatmap was constructed using R package *gplots*.

Selective analysis

The coefficient of nucleotide differentiation F_{ST} between the populations and Tajima's D and nucleotide diversity ($\theta\pi$) for each population were calculated by the Bio::PopGen::PopStats package in BioPerl (12). We calculated the log value of $\theta\pi$ ratios. The putative selected genes were screened from the overlap of the top 5% log-odds ratios of both $\theta\pi$ and F_{ST}. Functional enrichment of the candidate genes was performed by the ClueGO plugin of Cytoscape 3.2.1(19) using Symbol ID as input, and *Mus Musculus* was used as the background organism. After Bonferroni correction for multiple testing, $P < 0.05$ was considered to be statistically significant.

Supplementary Figures and Tables

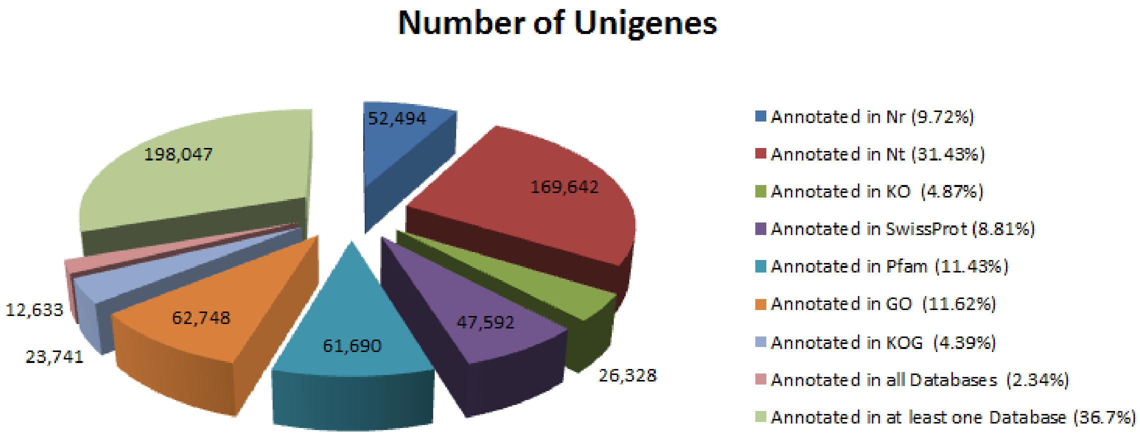

Fig. S1. Annotations of transcripts to different databases. These databases include Non-redundant (Nr), Nucleotide database (Nt), Swiss-prot, "KEGG (Kyoto Encyclopedia of Genes and Genomes)", and COG (Cluster of Orthologous Groups of proteins).

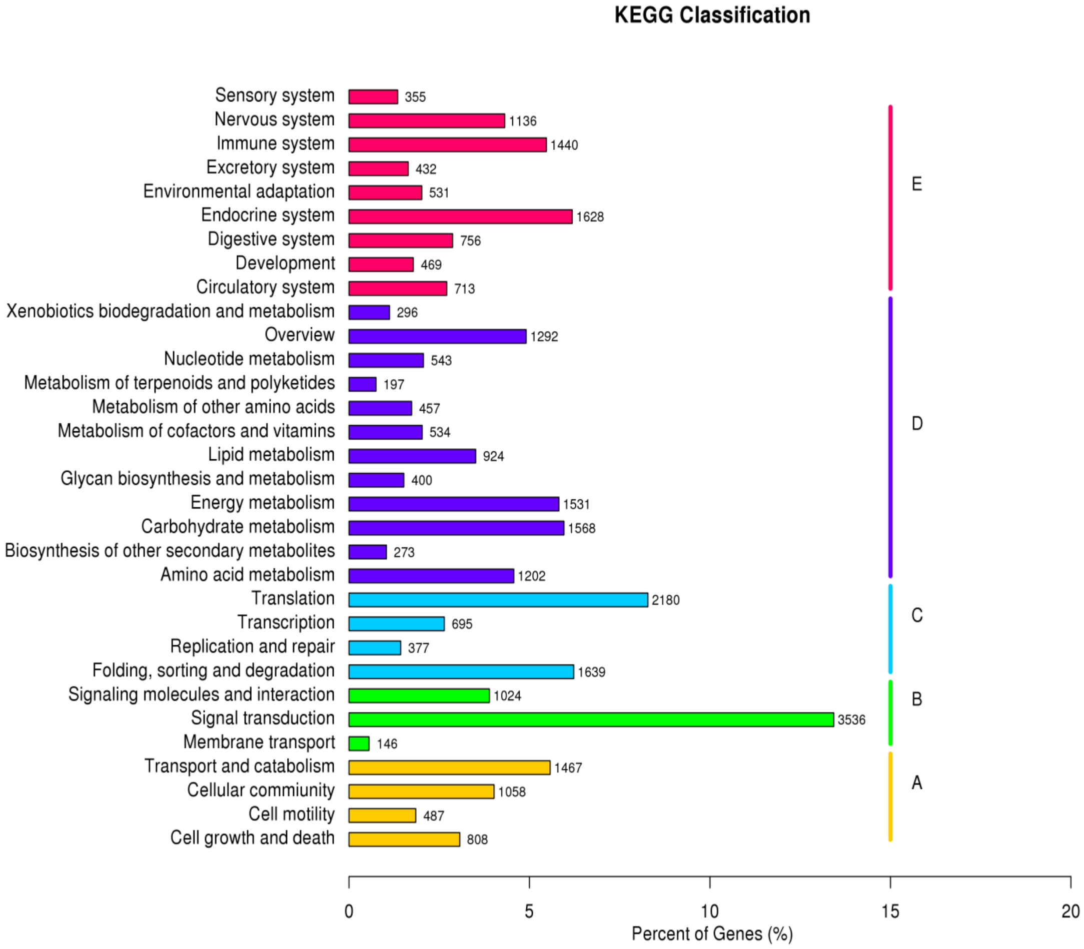

Fig. S2 Pathway assignment of unigenes by Kyoto Encyclopedia of Genes and Genomes (KEGG) classification. X-axis denote percentage of genes and Y-axis denote KEGG categories. A: Cellular Processes, B: Environmental Information Processing, C: Genetic Information Processing, D: Metabolism, E: Organismal Systems.

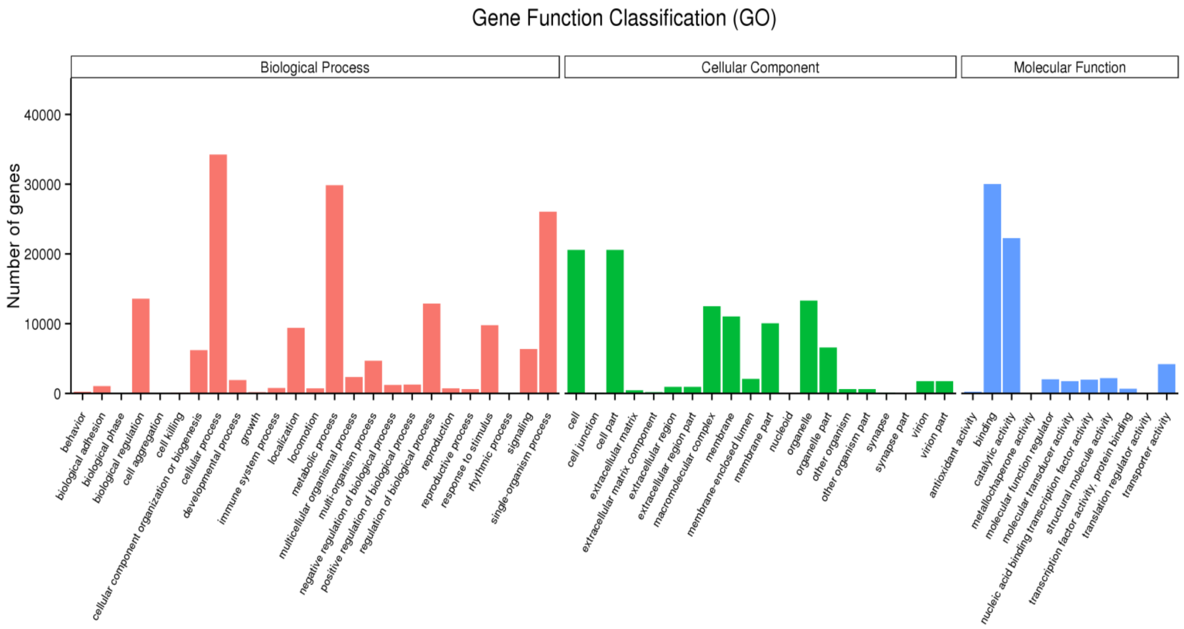

Fig. S3 Functional GO categories of the assembled transcriptome of *Acomys cahirinus* at EC I. The unigenes were annotated to biological process, molecular function, and cellular component. X-axis denote GO category and Y-axis were the number of genes.

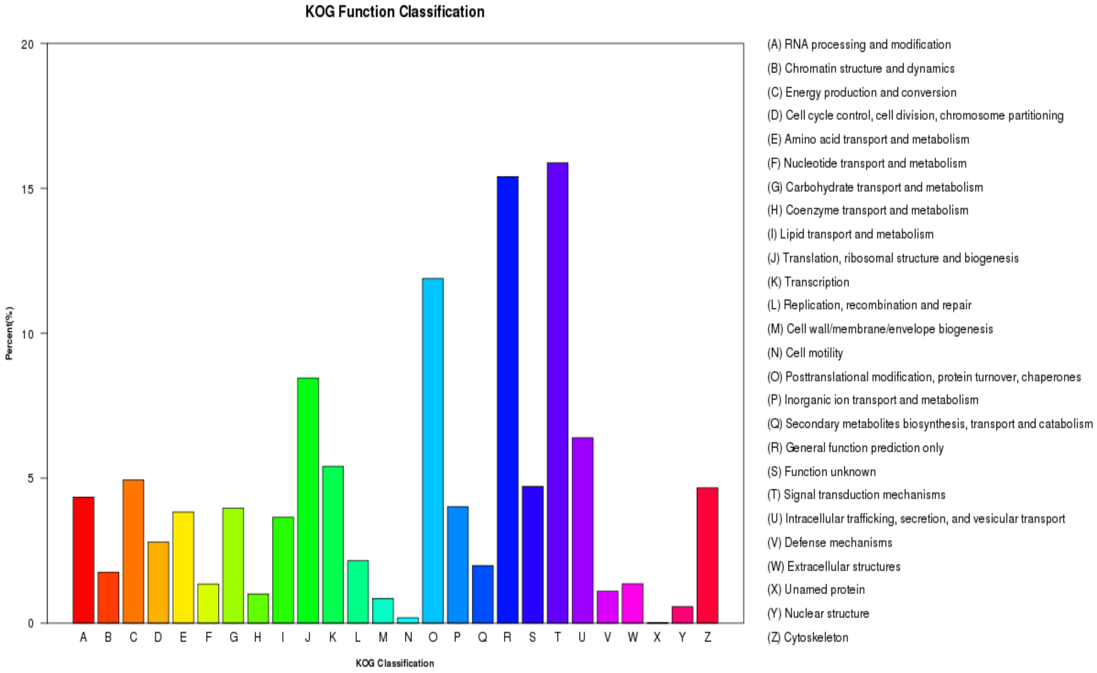

Fig. S4 EuKaryotic Orthologous Groups (KOG) annotation of putative classification of the transcriptome genes of *Acomys cahirinus* at EC I. X-axis denotes the KOG classification and Y-axis ware percentage of the genes.

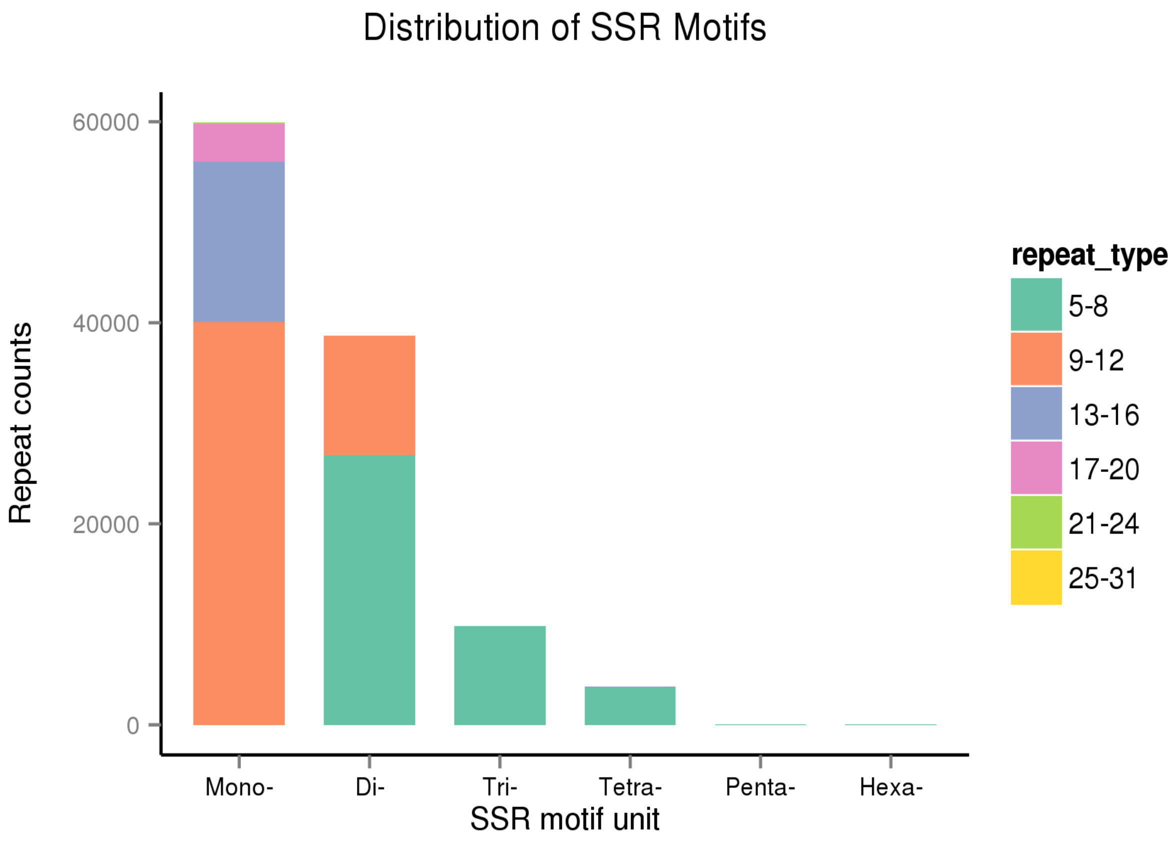

Fig. S5 Microsatellite motif distribution. Mono-, Di, Tri, Tetra-, Penta-, and Hexa- denote nucleotide number of the microsatellite unit. Repeat number was shown in different colors. X and Y axes denote SSR motif unit and repeat account, respectively.

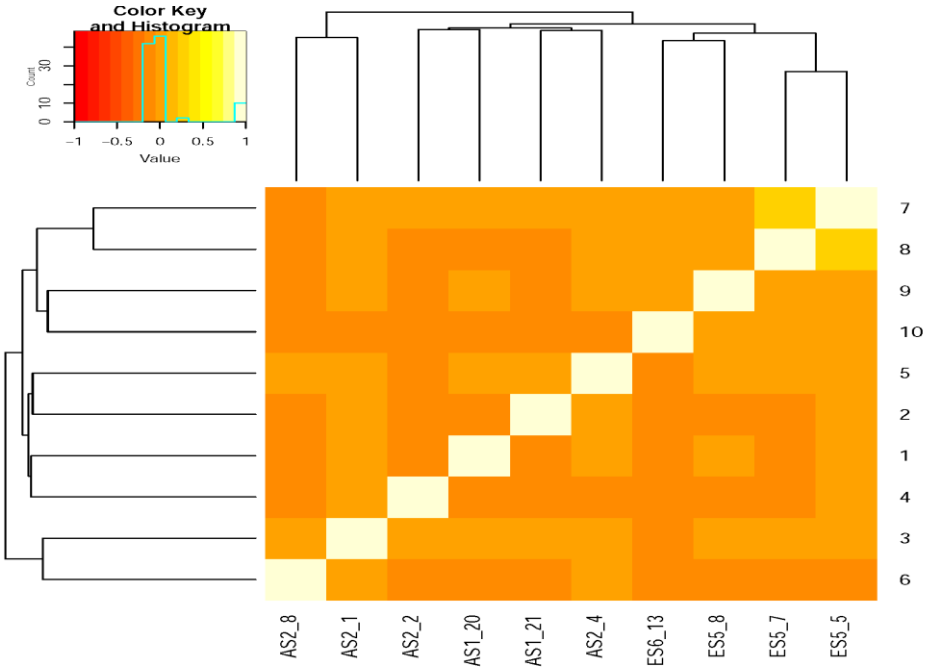

Fig. S6 Kinship of all animals from both AS and ES populations from Evolution Canyon I.

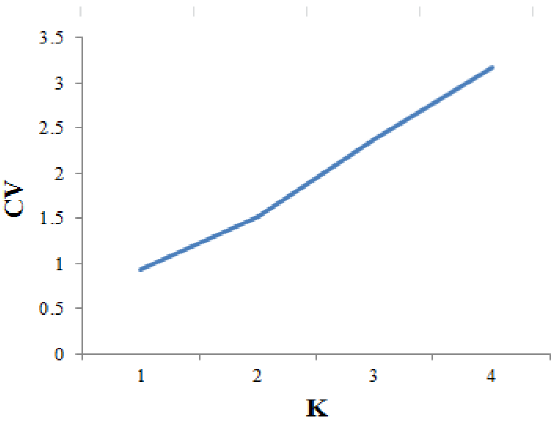

Fig. S7. Cross-validation error estimate and the K values for structure estimation

1. Nevo E (1995) Asian, African and European biota meet at Evolution Canyon Israel: local tests of global biodiversity and genetic diversity patterns. *Proc R Soc Lond B Biol Sci* 262(1364):149-155.
2. Pavl ćek T, Sharon D, Kravchenko V, Saaroni H, & Nevo E (2003) Microclimatic interslope differences underlying biodiversity contrasts in Evolution Canyon, Mt. Carmel, Israel. *Isr J Earth Sci* 52(1).
3. Gordon A & Hannon G (2010) Fastx-toolkit. *FASTQ/A short-reads preprocessing tools (http://hannonlab. cshl. edu/fastx_toolkit)*.
4. Andrews S (2010) FastQC: A quality control tool for high throughput sequence data. *Reference Source*.
5. Grabherr MG, *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnol* 29(7):644-652.
6. Li W & Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22(13):1658-1659.
7. Iseli C, Jongeneel CV, & Bucher P (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *ISMB*, pp 138-148.
8. Conesa A, *et al.* (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21(18):3674-3676.
9. McKenna A, *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20(9):1297-1303.
10. Li H & Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25(14):1754-1760.
11. Li H, *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* 25(16):2078-2079.
12. Stajich JE, *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res* 12(10):1611-1618.
13. Price AL, *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38(8):904-909.
14. Alexander DH, Novembre J, & Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 19(9):1655-1664.
15. Rousset F (2008) genepop’ 007: a complete re‐implementation of the genepop software for Windows and Linux. *Mol Ecol Resour* 8(1):103-106.
16. Paradis E, Claude J, & Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20(2):289-290.
17. Huson DH & Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 23(2):254-267.
18. Manichaikul A, *et al.* (2010) Robust relationship inference in genome-wide association studies. *Bioinformatics* 26(22):2867-2873.
19. Bindea G, *et al.* (2009) ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* 25(8):1091-1093.

Supplementary Appendix Suggested Readings

Forbes AA, Powell TH, Stelinski LL, Smith JJ, & Feder JL (2009) Sequential sympatric speciation across trophic levels. *Science* 323(5915):776-779.

Rolán-Alvarez E (2007) Sympatric speciation as a by-product of ecological adaptation in the Galician Littorina saxatilis hybrid zone. *J Mollus Stud* 73(1):1-10.

Doebeli M & Dieckmann U (2003) Speciation along environmental gradients. *Nature* 421(6920):259-264.

Nosil P (2008) Speciation with gene flow could be common. *Mol Ecol* 17(9):2103-2106.

Papadopulos AS, *et al.* (2011) Speciation with gene flow on Lord Howe Island. *Proc Natl Acad Sci USA* 108(32):13188-13193.

Martin CH (2013) Strong assortative mating by diet, color, size, and morphology but limited progress toward sympatric speciation in a classic example: Cameroon crater lake cichlids. *Evolution* 67(7):2114-2123.

Jones AG, Moore GI, Kvarnemo C, Walker D, & Avise JC (2003) Sympatric speciation as a consequence of male pregnancy in seahorses. *Proc Natl Acad Sci USA* 100(11):6598-6603.

Friesen V, *et al.* (2007) Sympatric speciation by allochrony in a seabird. *Proc Natl Acad Sci USA* 104(47):18589-18594.

Gavrilets S & Waxman D (2002) Sympatric speciation by sexual conflict. *Proc Natl Acad Sci USA* 99(16):10533-10538.

Higashi M, Takimoto G, & Yamamura N (1999) Sympatric speciation by sexual selection. *Nature* 402(6761):523-526.

Van Leuven JT, Meister RC, Simon C, & McCutcheon JP (2014) Sympatric speciation in a bacterial endosymbiont results in two genomes with the functionality of one. *Cell* 158(6):1270-1280.

Crow KD, Munehara H, & Bernardi G (2010) Sympatric speciation in a genus of marine reef fishes. *Mol Ecol* 19(10):2089-2105.

Via S (2001) Sympatric speciation in animals: the ugly duckling grows up. *Trends Ecol Evol* 16(7):381-390.

Barluenga M, Stölting KN, Salzburger W, Muschick M, & Meyer A (2006) Sympatric speciation in Nicaraguan crater lake cichlid fish. *Nature* 439(7077):719-723.

Savolainen V, *et al.* (2006) Sympatric speciation in palms on an oceanic island. *Nature* 441(7090):210-213.

Berlocher SH & Feder JL (2002) Sympatric speciation in phytophagous insects: moving beyond controversy? *Annu Rev Entomol* 47(1):773-815.

Bird CE, Fernandez-Silva I, Skillings DJ, & Toonen RJ (2012) Sympatric speciation in the post “modern synthesis” era of evolutionary biology. *Evol Biol* 39(2):158-180.

Bolnick DI (2011) Sympatric speciation in threespine stickleback: why not? *Int J Ecol* 2011.

Bolnick DI & Fitzpatrick BM (2007) Sympatric speciation: models and empirical evidence. *Annu Rev Ecol Evol*:459-487.

Schliewen UK, Tautz D, & Pääbo S (1994) Sympatric speciation suggested by monophyly of crater lake cichlids. *Nature* 368(6472):629-632.

Friedman J, Alm EJ, & Shapiro BJ (2013) Sympatric speciation: when is it possible in bacteria? *PLoS ONE* 8(1):e53539.

Jiggins CD (2006) Sympatric speciation: why the controversy? *Curr Biol* 16(9):R333-R334.

Feder JL, Egan SP, & Nosil P (2012) The genomics of speciation-with-gene-flow. *Trends Genet* 28(7):342-350.

Martin CH (2012) Weak disruptive selection and incomplete phenotypic divergence in two classic examples of sympatric speciation: Cameroon Crater Lake cichlids. *Am Nat* 180(4):E90-E109.

Fitzpatrick BM, Fordyce J, & Gavrilets S (2008) What, if anything, is sympatric speciation? *J Evol Biol* 21(6):1452-1459.

Michel AP, *et al.* (2010) Widespread genomic divergence during sympatric speciation. *Proc Natl Acad Sci USA* 107(21):9724-9729.

Soria-Carrasco V, *et al.* (2014) Stick insect genomes reveal natural selection’s role in parallel speciation. *Science* 344(6185):738-742.

Débarre F (2012) Refining the conditions for sympatric ecological speciation. *J Evol Biol* 25(12):2651-2660.

Johannesson K (2001) Parallel speciation: a key to sympatric divergence. *Trends Ecol Evol* 16(3):148-153.

Dieckmann U & Doebeli M (1999) On the origin of species by sympatric speciation. *Nature* 400(6742):354-357.

Kondrashov AS & Kondrashov FA (1999) Interactions among quantitative traits in the course of sympatric speciation. *Nature* 400(6742):351-354.

Wilson AB, Noack–Kunnmann K, & Meyer A (2000) Incipient speciation in sympatric Nicaraguan crater lake cichlid fishes: sexual selection versus ecological diversification. *Phil Trans R Soc B* 267(1458):2133-2141.

Gourbiere S (2004) How do natural and sexual selection contribute to sympatric speciation? *J Evol Biol* 17(6):1297-1309.

Papadopulos AS, *et al.* (2014) Evaluation of genetic isolation within an island flora reveals unusually widespread local adaptation and supports sympatric speciation. *Phil Trans R Soc B* 369(1648):20130342.

Boomsma JJ & Nash DR (2014) Evolution: sympatric speciation the eusocial way. *Curr Biol* 24(17):R798-R800.

Les DH, *et al.* (2015) Through thick and thin: Cryptic sympatric speciation in the submersed genus Najas (*Hydrocharitaceae*). *Mol Phylogenet Evol* 82:15-30.