FAST TRACK

Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing

J. CRISTOBAL VERA,*¶ CHRISTOPHER W. WHEAT,*†¶ HOWARD W. FESCEMYER,*
MIKKO J. FRILANDER,‡ DOUGLAS L. CRAWFORD,§ ILKKA HANSKI† and JAMES H. MARDEN*
*Department of Biology, 208 Mueller Laboratory, Pennsylvania State University, University Park, PA 16802, USA,
†Department of Biological and Environmental Sciences, University of Helsinki, Viikinkaari 1 PL 65, 00014 Helsinki, Finland,
‡Institute of Biotechnology, University of Helsinki, Viikinkaari 9 PL 56, 00014 Helsinki, Finland, §Rosenstiel School of Marine and
Atmospheric Sciences, University of Miami, 4600 Rickenbacker Causeway, Miami, FL 33149, USA*

Abstract

We present a *de novo* assembly of a eukaryote transcriptome using 454 pyrosequencing data. The Glanville fritillary butterfly (*Melitaea cinxia*; Lepidoptera: Nymphalidae) is a prominent species in population biology but had no previous genomic data. Sequencing runs using two normalized complementary DNA collections from a genetically diverse pool of larvae, pupae, and adults yielded 608 053 expressed sequence tags (mean length = 110 nucleotides), which assembled into 48 354 contigs (sets of overlapping DNA segments) and 59 943 singletons. BLAST comparisons confirmed the accuracy of the sequencing and assembly, and indicated the presence of *c*. 9000 unique genes, along with > 6000 additional microarray-confirmed unannotated contigs. Average depth of coverage was 6.5-fold for the longest 4800 contigs (348–2849 bp in length), sufficient for detecting large numbers of single nucleotide polymorphisms. Oligonucleotide microarray probes designed from the assembled sequences showed highly repeatable hybridization intensity and revealed biological differences among individuals. We conclude that 454 sequencing, when performed to provide sufficient coverage depth, allows *de novo* transcriptome assembly and a fast, cost-effective, and reliable method for development of functional genomic tools for nonmodel species. This development narrows the gap between approaches based on model organisms with rich genetic resources vs. species that are most tractable for ecological and evolutionary studies.

Keywords: bioinformatics, biotechnology, functional genomics, metapopulation, polymorphism, transcriptomics

Received 8 September 2007; revision accepted 5 December 2007

Introduction

Whole genome or transcriptome sequencing enables functional genomic studies based on global gene expression, single nucleotide polymorphism (SNP) surveys, quantitative trait loci (QTL) studies, genomic scans of diversity, and so forth (Rudd 2003; Bouck & Vision 2007; Nagaraj *et al.* 2007). However, most species studied for ecological or physiological traits do not have genomic or expressed sequence tag (EST) data available. For nonmodel species, the method of

obtaining transcriptome data has been complementary DNA (cDNA) library construction, repeated rounds of normalization/subtraction and Sanger EST sequencing, and often cDNA microarray construction (Whitfield *et al.* 2002; Mita *et al.* 2003; Rudd 2003; Paschall *et al.* 2004; Papanicolaou *et al.* 2005; Beldade *et al.* 2006). Those proven methods can potentially be improved upon by next-generation approaches (Hudson 2007) that reduce costs, labour, errors associated with clone mishandling, and that recover missing or rare transcripts or those that are unstable when cloned into bacteria (Weber *et al.* 2007).

Here, we demonstrate the utility of recently developed sequencing technology for rapid and accurate transcriptome characterization. Our subject species, the Glanville fritillary

Correspondence: J. Cris Vera, Fax: 814-865-9131; E-mail: jcv128@psu.edu
¶J.C. Vera and C.W. Wheat contributed equally to this work.

butterfly (*Melitaea cinxia*), has been examined extensively from an ecological and population biology perspective (Ehrlich & Hanski 2004) and is polymorphic for key traits that affect fitness in ways that are dependent on habitat patch size, connectivity, and population history (Hanski *et al*. 2004; Haag *et al*. 2005; Hanski & Saccheri 2006; Hanski *et al*. 2006; Saastamoinen 2007a, b). As is typical for species studied primarily in their free-living state, there are no genomic resources available for mechanistic dissection of these ecologically interesting traits. We report here how 454 pyrosequencing and assembly of short sequence reads can be used to generate functional genomic tools for this and potentially any species.

Massively parallel 454 pyrosequencing has become a feasible method for increasing sequencing depth and coverage while reducing time, labour, and cost (Margulies *et al*. 2005; Moore *et al*. 2006; Wicker *et al*. 2006; Huse *et al*. 2007; Weber *et al*. 2007). Typical output from a 4.5-h run of the original GS20 sequencer is 20–30 Mb, comprising roughly 300 000 sequence reads averaging *c*. 100 bp. Sequencing error levels are low (< 1%), arising primarily because of homopolymer runs (Margulies *et al*. 2005; Moore *et al*. 2006; Huse *et al*. 2007), but these errors tend to be resolved in cases where there is sufficient coverage depth to allow assembly of overlapping reads. For this reason, 454 sequences can be more accurate than traditional Sanger sequences (Goldberg *et al*. 2006; Moore *et al*. 2006; Wicker *et al*. 2006). However, the short sequence reads make assembly of overlapping sequences problematic (Trombetti *et al*. 2007). The challenge is greatest when genomic data are not available to aid assembly, when there is polymorphism, and when the data are cDNA sequences that contain variation created by alternative splicing and vary widely in transcript abundance, making coverage uneven.

Previous studies have used 454 pyrosequencing and assembly with polymerase chain reaction (PCR) amplicons, bacterial artificial chromosomes (BAC), genomic, mitochondrial, and plastid DNA (e.g. Bainbridge *et al*. 2006; Goldberg *et al*. 2006; Moore *et al*. 2006; Poinar *et al*. 2006; Wicker *et al*. 2006). So far, published reports of 454 pyrosequencing of transcriptomes has been restricted to model species where genomic or extensive Sanger EST data provided an assembly reference (Bainbridge *et al*. 2006; Cheung *et al*. 2006; Emrich *et al*. 2007; Weber *et al*. 2007). Two studies (Cheung *et al*. 2006; Weber *et al*. 2007) that used genome or Sanger EST sequences for mapping and annotation of 454 ESTs (Cheung *et al*. 2006; Weber *et al*. 2007) were not able to also accomplish *de novo* assembly of their 454 ESTs (e.g. only two contigs out of 184 599 reached 500 bp; Cheung *et al*. 2006). A transcriptome study using 454 ESTs from a wasp accomplished annotation by alignment with the honeybee genome, and used sequence data to design primers for quantitative reverse transcription–PCR (qRT–PCR), but did not report an assembly (Toth *et al*.

2007). Here, we test the suitability of 454 data for *de novo* assembly and annotation of expressed genes of a eukaryote, and show the utility of these data for designing oligonucleotide microarrays. The result establishes a workflow from RNA to gene assembly and large-scale functional genomics tools.

Materials and methods

RNA was isolated from larvae, pupae, and adults of *Melitaea cinxia* (*c*. 80 individuals from eight families) collected from old and new populations across the main Åland island located in the Baltic between Finland and Sweden. As appropriate for SNP discovery, this sample is likely to be representative of the genetic diversity of the overall Åland metapopulation. RNA was extracted from whole bodies or body segments. We combined equivalent amounts of RNA into two pools, (i) larvae plus pupae, and (ii) adult, which together contained 16.1 mg of total RNA. Complementary DNA was synthesized from the two RNA pools and normalized (Zhu *et al*. 2001; Shagin *et al*. 2002; Zhulidov *et al*. 2005; Evrogen: www.evrogen.com) in order to prevent over-representation of the most common transcripts (Fig. S1, Supplementary material). Approximately 1 μg of double-stranded cDNA from each of two normalized cDNA pools was sequenced on a Roche GS20 sequencer using methods previously described (Margulies *et al*. 2005; Poinar *et al*. 2006).

Initial quality filtering of the 454 ESTs was performed at the machine level before base calling. The 454 EST sequences were screened for primer sequence using a custom perl script (SmartScreener). Screened ESTs and their quality scores were then assembled using commercially available software (SEQMAN PRO 7.1). Before assembly, quality filtering was performed using default quality parameters for nontrace sequences. Additional assembly parameters were those identified as optimal by the manufacturer for large 454 data sets (Table S7, Supplementary material). The assembler utilized a combined quality weighting (shallow coverage) and simple majority (deep coverage) for consensus base calling after assembly alignment. Files containing our sequences and their quality scores are available from the National Center for Biotechnology Information (NCBI) Short Read Archive, accession SRA000207.

In order to compare 454 sequences against Sanger sequences, we generated plasmid cDNA libraries. We used 5 μg of larval messenger RNA (mRNA) for cDNA synthesis, essentially following a previously described protocol (Paschall *et al*. 2004). The cDNAs were unidirectionally cloned into a vector and sequenced from the 5′-end. Sanger ESTs were assembled using SEQMAN PRO 7.1 with the default parameters for Sanger data and quality trimming set to high stringency. All Sanger ESTs were screened for vector sequence.

All 454 and Sanger sequences (contigs and singletons) were aligned (NCBI BLASTX) vs. three expanded subvolumes of the Uniprot 9.2 annotated protein database (http://www.expasy.uniprot.org/database/download.shtml). Uniprot BLAST results were passed through a custom pipeline to create annotated, tab-delimited tables, which included available information on taxonomy, key words, gene function, tissue specificity, and gene ontology (GO) terms.

Additional alignments were performed as follows. The *in silico* predicted *Bombyx mori* protein set (Wang *et al*. 2005; http://silkworm.genomics.org.cn/), *Drosophila melanogaster* unigenes (Flybase: http://flybase.bio.indiana.edu/), and *Heliconius erato* clustered ESTs (Papanicolaou *et al*. 2005) were used to estimate coverage of the *M. cinxia* transcriptome (NCBI BLASTX), whereas the Butterflybase version 2.91 (Papanicolaou *et al*. 2005) combined EST set was used primarily as a catch-all for otherwise uncharacterized sequences (NCBI TBLASTX). Finally, both the Sanger contigs and the 454 contigs were aligned (NCBI BLASTN, no filters) against the combined 454 contig/singleton data set in order to analyse the quality of the assembly. All sequences, BLAST results, and annotation tables were stored in a normalized MySQL database.

The sense strand of contigs and singletons (both 454 and Sanger) that generated acceptable BLAST hits (bitscore > 45) to either *B. mori*-predicted proteins or to protein databases, along with both strands for those that hit Butterflybase ESTs (for which strand orientation is uncertain), were submitted to Agilent's eArray web tool (http://earray.chem.agilent.com/earray/) to generate oligonucleotide (60-mer) microarray probe sequences. Six different probes were generated for each submitted sequence. In addition, one probe was generated for the two complementary strand sequences of each contig that was not identified in BLAST searches; this allowed strand orientation of the unannotated contigs to be inferred from the difference in relative hybridization intensity of the two complementary probes.

All probes (*N* = 207 149) were printed along with Agilent positive and negative controls on a 244K-feature Agilent microarray. Fluorescent-labelled complementary RNA (cRNA) (Cy3) was synthesized from an even mixture of the RNA pools used originally for the normalization and 454 sequencing; these were hybridized to the array using standard Agilent reagent kits and protocols. Three replicates of the best-performing probe from 13 780 annotated contigs and singletons were printed on 44K-feature arrays, along with a small number of probes for Sanger EST's and sense-strand inferred probes for otherwise unannotated contigs. Two arrays were hybridized with fluorescent-labelled antisense RNA (aRNA) from two individual butterfly abdomens. Repeatability was assessed by comparing mean intensity of the replicated probes for one labelled aRNA sample across two arrays. Array probes, layout, and access to purchase this microarray from Agilent can be obtained from the senior author (J.H.M.).

A more detailed description of Materials and Methods is contained in Appendix S1, Supplementary material.

Results

Sequencing and assembly

A total of 608 053 ESTs averaging 110 bp (see Table S1, Supplementary material) were obtained from two 454 sequencing runs. Of these, 518 079 exceeded our minimal quality standards (SMART primer filtering; threshold of 50-bp windows with mean quality score > 12) and entered the assembly. Eighty-eight per cent (458 136) of the high quality ESTs formed 48 354 contigs of average length of 197 bp (median = 134; SD = 207), with an average depth of coverage of 2.97 sequences per nucleotide position (Fig. 1). The longest 10% (*n* = 4835; 348–2849 bp in length) had an average depth of coverage of 6.5. Thus, even though the average contig was short, we obtained thousands of long and deeply covered contigs. The remaining 59 943 high-quality ESTs were retained as singletons (coverage depth = 1), for a total of 108 297 unigenes.

An additional 3888 sequence reads were obtained from *Melitaea cinxia* cDNA libraries using traditional Sanger sequencing techniques. Forty-four per cent (1711) of the vector-screened and quality-trimmed Sanger ESTs assembled into 364 contigs with average length of 574 bp and average coverage depth of 3.0 (see Table S2, Supplementary material). There were 1297 Sanger ESTs that remained singletons. These data were used to test the quality of the 454 sequencing and assembly.

Quality and performance of the 454 assembly

Ideally, accurate and effective assembly of the 454 sequence reads would create assembled sequences with at most a few short alignments to other 454 sequences (e.g. conserved motifs). We tested this with a BLASTN alignment of all 48K contigs against themselves and the singletons. Eight per cent of the contigs (3897) had BLAST hits (bitscore > 45) with 100% identity with other contigs (*N* = 1749) and singletons (*N* = 2660), but in no case did these alignments extend over the entire length of either the BLAST subject or query. These perfect match alignments averaged 35 and 36 nucleotides in length for contig and singleton hits, respectively, averaging 18% (median = 16%; maximum = 82%) of the length of the query contig. Nearly all of these contigs (3892) had BLASTX hits (bitscore > 45) against *Bombyx mori* proteins (Wang *et al*. 2005), yet only 220 (6%) of those BLAST-paired contigs had best BLASTX hits to the same *B. mori* protein, and only 69 (2%) of those pairs also had best BLASTX hits to the same protein in Uniprot. Thus, these
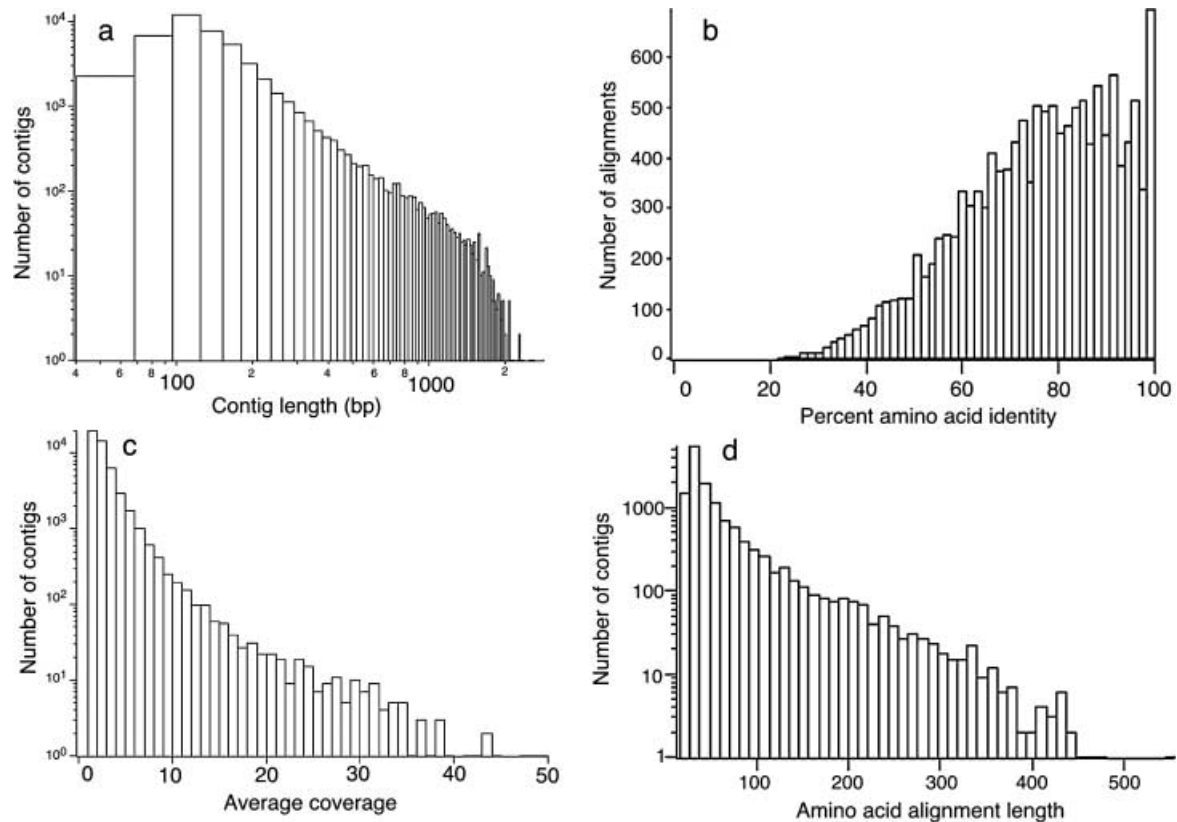
Fig. 1 Characteristics of assembled *Melitaea cinxia* 454 contigs, and BLASTX alignments against *Bombyx mori*. (a,b) Length and coverage of contigs. Variable bar width in plot A is an artefact of the logarithmic horizontal axis; there is no information content in the area of these bars. Note the logarithmic vertical axes; the majority of contigs were small and had low coverage depth. (c,d) Frequency distributions of percent identity and deduced amino acid alignment length for all blast hits (bitscore > 45) of *M. cinxia* contigs and singletons (N = 13 653) to 6289 *B. mori* predicted proteins.

gene regions that produced 100% identity alignments were primarily conserved motifs in different genes, or in some cases variants of single genes created by alternative splicing; our assembly appropriately partitioned these into different contigs.

The 454 assembly was further assessed by comparison to Sanger ESTs derived from an *M. cinxia* cDNA library. Of 813 quality-trimmed Sanger contigs and singletons filtered for length and annotation (> 200nt and with BLAST hits to metazoan genes), 749 (92%) had strong BLAST hits to 454 sequences (Table S3, Supplementary material). On average, 74% of the length of these Sanger sequences was aligned to at least 1× depth by 454 sequence (median = 90%). Nucleotide alignments of Sanger vs. 454 sequences were 97% identical for all alignments and 99% for each Sanger sequence's best BLAST alignment (by bitscore; Fig. S2, Supplementary material). The average number of gaps for alignments involving 454 contigs was 0.63 (median = 0; Table S3) over an average length of 200 nucleotides. This gap rate of about three per thousand aligned bases was less than the seven gaps per thousand aligned bases for 454 singletons, which because of 1× coverage contain a

higher frequency of homopolymer run errors. These are overestimates of the true 454 base and gap error rates, as they include base mismatches caused by polymorphism, gaps created by alternative splicing, and alignment with less accurate end regions of Sanger sequences.

A number of factors affect *de novo* assembly of genes from 454 sequence data (Figs S3 and S4, Supplementary material), especially coverage depth. The effect of coverage depth is apparent in a BLASTX analysis of 454 contigs to *B. mori* predicted proteins (Wang *et al.* 2005; the only fully sequenced lepidopteran genome). The ratio of the length of individual contigs to the length of their *B. mori* orthologue coding region increased with the depth of coverage of the contig and reached an asymptote near unity (i.e. likely full length gene assemblies) at average coverage > 10, for all but the longest genes (Fig. 2a). From the distribution of average coverage depth for these contigs (median = 2.78), we infer that a fourfold increase in sequencing effort of our cDNA pool would be required to achieve an average coverage depth of 11 for about half of the contigs, at which point *c*. 50% of the genes would likely assemble to full length. Greater coverage depth in the 454 assembly increased
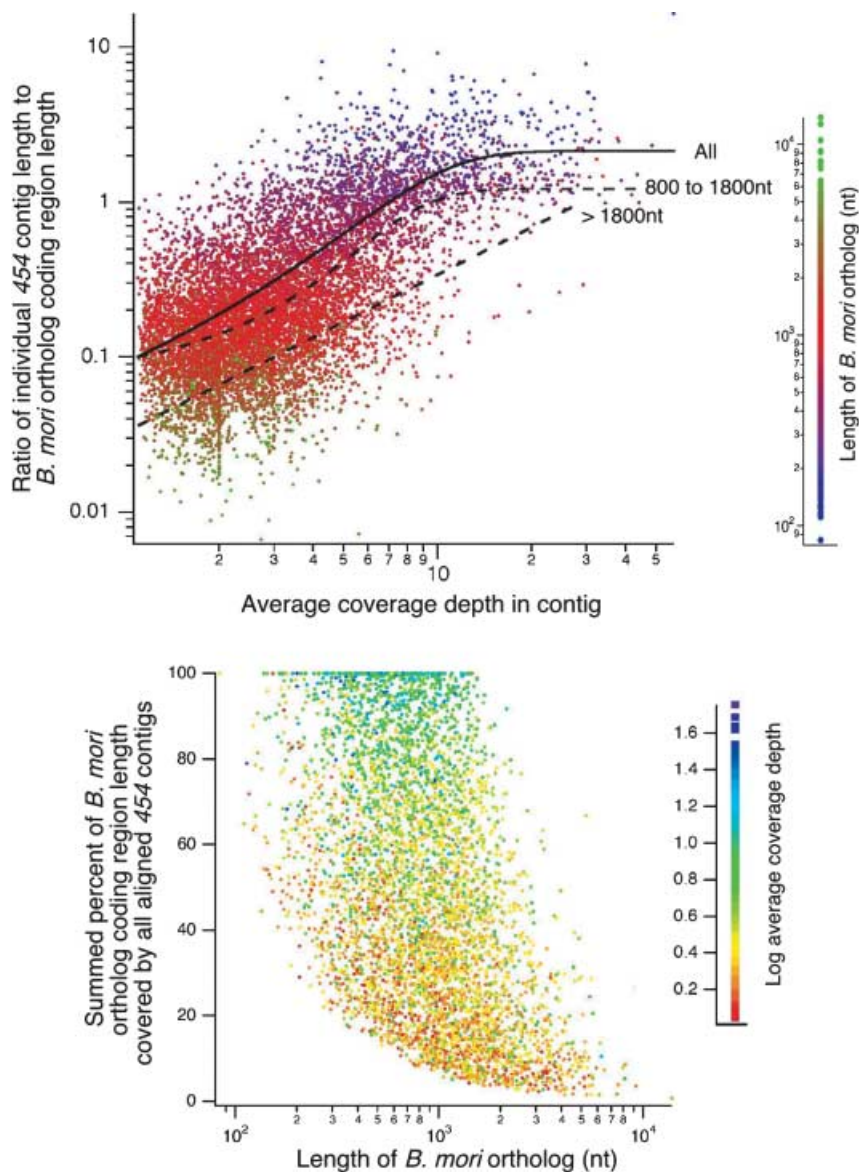
Fig. 2 Comparison of *Melitaea cinxia* 454 contig length to the coding region length of *Bombyx mori* orthologous genes. Orthologues were identified by BLASTX (bitscore > 45). (a) Ratio of 454 contig to *B. mori* orthologue length in relation to coverage depth (average number of ESTs per nucleotide site within a contig). Ratios greater than unity are caused primarily by the presence in 454 contigs of untranslated regions that extend beyond the coding region. Solid curve shows sigmoid fit for all *B. mori* genes; dashed curve shows fit for *B. mori* genes 800–1800 bp in length; line shows fit for those greater than 1800 bp length. Where average coverage depth was = 10, nearly all 454 contigs were as long or longer than the coding region of their putative *B. mori* orthologue, suggesting full-length assembly. (b) Ratio of *B. mori* orthologue length to the summed nonoverlapping alignment length of *M. cinxia* sequences that had a best BLAST hit (bitscore > 45) to that *B. mori* gene (i.e. all putative fragments of individual genes).

the total percentage length alignment by all contigs across their respective *B. mori* orthologue, in a manner that depended also on *B. mori* gene length (Fig. 2b). Total coverage was best in 573 cases where *B. mori* proteins had at least 90% of their coding region aligned to one or more *M. cinxia* contig.

Transcriptome coverage breadth

Among the 108 297 contigs and singletons, 19 383 (18%) had hits exceeding our threshold (bitscore > 45, *e*-value generally < 1^{-5}) in at least one of three BLAST searches: BLASTX against predicted *B. mori* proteins (Wang *et al.* 2005); TBLASTX against clustered lepidopteran ESTs (Butterflybase; Papanicolaou *et al.* 2005), or BLASTX against Uniprot proteins (Boeckman *et al.* 2003; Fig. 3a). Excluding sequences with best BLAST hits to non-metazoa, 71% (13 653) were similar

to 6289 unique *B. mori* proteins, with an average best-hit amino acid identity of 73% (SD = 16, Fig. 1c, d; Table S4, Supplementary material). In addition, sequences with no *B. mori* match had best BLAST hits (bitscore > 45) to 1748 metazoan Uniprot proteins with unique gene descriptions, and an additional 1274 unique EST clusters in Butterflybase, indicating the presence of about 9.3K (6289 + 1748 + 1274) *M. cinxia* unigenes. Comparisons against *Drosophila melanogaster* (Diptera) proteins and *Heliconius erato* (Lepidoptera) clustered ESTs produced similar results (Fig. 3b). This minimal estimate of 9.3K annotated genes is nearly two-thirds of the number of genes in *D. melanogaster* (13 379; Adams *et al.* 2000), about half of the number of genes in *B. mori* (predicted genes = 18 510; Xia *et al.* 2004), and exceeds what has been found by all of the Sanger sequencing of Lepidoptera ESTs other than *B. mori*. The
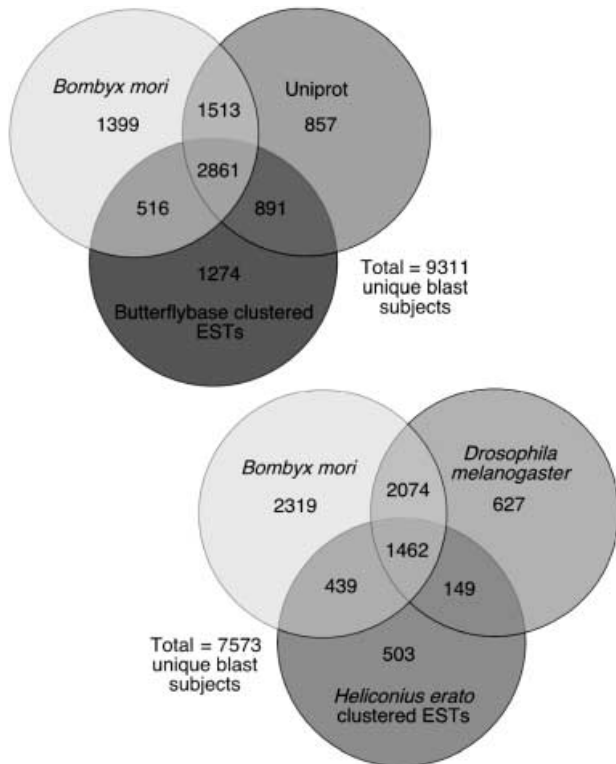
Fig. 3 Venn diagram showing distribution of similarity search results. Numbers are the sum of unique *Bombyx mori* proteins, unique *Drosophila melanogaster* or Uniprot proteins, and unique Butterflybase (Lepidoptera species excluding *B. mori*) or *Heliconius erato* EST clusters that had best blast hits (> 45 bitscore) to a 454 unigene. Priority hierarchy for determining uniqueness of proteins or EST clusters in overlap regions (i.e. cases where multiple 454 unigenes aligned to a unigene shared between the respective databases) was *B. mori* protein > Uniprot protein or *D. melanogaster* protein > EST cluster name. Sequences with best blast hits to non-metazoan proteins are excluded from these counts.

two lepidoptera (*H. erato*, *Spodoptera frugiperda*; Butterflybase) with the most EST samples contain approximately 6–7K putative protein-coding genes. A BLASTX search using butterfly ESTs publicly available (Butterflybase; N = 28 664; *Bicyclus anynana*, *H. erato*, *H. melpomene*, *Papilio dardanus*, *P. xuthus*) against the *B. mori* predicted proteins produced 15 920 hits (bitscore > 45) to 5085 *B. mori* proteins (24% less than our *M. cinxia* 454 assembly).

The percentage of *M. cinxia* 454 unigenes that were annotated (18%) is similar to that observed in other large assembled EST data sets of nonmammalian species (Mita *et al*. 2003; Paschall *et al*. 2004), suggesting that the unannotated contigs and singletons could represent a substantial fraction of the *M. cinxia*-specific transcriptome (as previously observed in butterflies; Beldade *et al*. 2006). Using an Agilent microarry (60-mer probes) to determine if these unannotated sequences are expressed mRNAs, we found that the distribution of signal intensities was quite

similar among probes for annotated and unannotated genes (Fig. S5, Supplementary material). Probe signal intensity for 6156 unannotated contigs exceeded the 99.5th percentile of negative control probes, and this number is likely to have been much higher on a more effectively labelled array, as this one had no saturated probes and therefore was not effective for detecting rarer transcripts. There was a clear effect of probe orientation (average 10-fold difference) when one probe orientation produced a strong signal (Fig. S5d). This further validates transcript presence of non-annotated contigs because only probes complementary to valid mRNA sequences are likely to produce detectable signals on an Agilent oligonucleotide (60-mer) microarray (Hughes *et al*. 2001).

Functional annotation

Further characterization of 454 ESTs was carried out with respect to the curated protein database of Uniprot/Swiss Prot (Boeckman *et al*. 2003), and functionally annotated genes in *D. melanogaster* (Flybase: http://flybase.bio. indiana.edu/). Using these databases, each 454 contig or singleton was assigned a gene description and a GO classification based on the 'best hit' BLASTX search (bitscore > 45, *e*-value generally < 1⁻⁵), using the 'inferred from sequence similarity' (ISS) level of evidence (Ashburner *et al*. 2000). Biological processes make up the GO designation of the majority of Uniprot proteins (3860 counts, 64%), followed by molecular function (1944, 32%) and cellular components (243, 4%). Distribution of gene annotations was similar for *D. melanogaster* (biological processes = 2560 counts, 54%; molecular function = 2086, 44%; cellular components = 127, 3%; Table S5, Supplementary material).

SNP discovery

The assembled 454 contigs provide a rich data source for discovery of common SNPs. Visual inspection of our 10 longest Uniprot-annotated contigs, using SNP identification criteria of at least 2× occurrence of the minority allele, at sites covered by at least five ESTs (mean coverage depth = 10.3) and away from homopolymer runs revealed 6.7 SNPs per thousand base pairs (N = 232), which matches closely the SNP density within gene-coding regions in mosquitoes (Wondji *et al*. 2007) but is less than that found in *M. cinxia* genomic DNA (Orsini *et al*. 2007). To detect SNPs of known codon position over a larger scale, we used custom scripts and database searching to find occurrences of single polymorphic sites within BLAST-annotated regions of contigs where there were no gaps, and where the minority allele was present in at least 25% of ESTs at sites with at least 6× coverage. There were 751 of these SNPs in 355 contigs (59 747 total sites within aligned regions; 12.6 SNPs per thousand base pairs), comprising

Fig. 4 Alternative splicing near the 3' end of the previously characterized *Melitaea cinxia* muscle gene *troponin-i*. (a) Conjoining of ESTs (red, green, and yellow boxes) representing partial nucleotide sequences of three alternatively spliced mutually exclusive versions of exon six in 454 assembly alignment. Note the degenerate nucleotide string in the consensus sequence (blue box). (b) Amino acid alignment of the same three alternatively spliced exons (red, green, and yellow exons) with *Bombyx mori* sequence for comparison (bottom yellow). The top alternative exon, A1a (purple), was recovered full length in a separate 454 contig. There was one additional, much shorter 454 contig partially spanning the alternatively spliced region representing most of TnI_A3a (green). Reassembly of all of the *troponin-i* 454 ESTs using more stringent assembly parameters produced four contigs spanning this region with full (A1a and A4a) and partial (A2a and A3a) coverage of the four versions of exon six. Full characterizations of three of the alternative exons (A2a, A3A, and A4a) had been achieved in previous experiments (J. Vera, unpublished) using standard PCR, RACE, and cloning methods and were needed to unravel the complexity of the affect of alternative splicing on these 454 contigs.

602 third position sites (likely to be synonymous), 79 first, and 70 second position sites. Thus, we identified 149 sites that are good candidates for amino acid polymorphisms within conserved regions of annotated genes.

Alternative splicing effects on assembly

Alternative splicing (AS) increases protein variation and complexity in eukaryotes (Marden 2006; Yasukochi *et al.* 2006; Auboeuf *et al.* 2007; Kim *et al.* 2007; Pajares *et al.* 2007) and poses difficulties for *de novo* assembly of short gene sequences. The only published study of AS in 454 sequences relied on an extensive Sanger EST database (human) to assess known and novel splicing (Bainbridge *et al.* 2006).

We assessed the effect of AS on our assembly by characterizing two AS genes (*troponin-t* and *troponin-i*) in *M. cinxia* using standard PCR, rapid amplification of cDNA ends (RACE), and cloning methods (J. Vera & M. Frilander, unpublished). Both genes had contigs in the 454 assembly with deep coverage that incorporated ESTs with small alternative or missing exons (or pieces thereof; Fig. 4), resulting in a short string of degenerate nucleotides in the consensus sequence. When ESTs contained longer alternatively spliced regions, a separate, shorter contig was formed encompassing primarily the alternatively spliced exon(s). In other cases, the contig was split at the splice site, creating two or more nonoverlapping sequences. Thus, although AS did not render assembly impossible or unusable, these

characterized AS regions assembled in a complex way that required clone sequences, genomic data from a related model species, and human processing to decipher.

Performance of microarray probes

To a 244K-feature microarray, we hybridized a 50:50 mix of labelled RNA from the same pools that were normalized and used for 454 sequencing. On this array were six different probes per BLAST-annotated contig and singleton, along with two complementary probes (i.e. from the obtained sequence and its complement) for non-annotated contigs. Many thousands of these probes produced hybridization signal intensities above the range of negative control probes (Fig. S5). Probes derived from 454 sequences performed as well as those from Sanger sequences (Fig. S6, Supplementary material). We printed three replicates of the best-performing 454 probe, assayed above, for each of 13 780 annotated contigs and singletons (~9K putative unique loci; Fig. 3) on Agilent 4× 44K format slides, and hybridized these with amplified mRNA (aRNA) from the abdomen of individual butterflies. Mean spot intensity for a single butterfly across replicate arrays (Fig. 5a inset) revealed excellent repeatability ($r^2 = 0.98$). These results further validate sequence quality and assembly, as well as the probe design method.

Metatrascriptomics and detection of an intracellular parasite

The cDNA for the 454 sequences came from whole insects whose surface and gut contained symbionts, parasites, and pathogens, some of which appeared in our annotated sequences. BLASTX analysis revealed 618 sequences with best hits (bitscore > 45) to non-metazoa (Table S6, Supplementary material). Individual non-metazoan taxa had relatively few hits, except for microsporidia ($N = 452$ hits). Selection of polyadenylated mRNAs during the formation of the cDNA pool for sequencing probably explains the relatively small number of bacterial sequences.

These microbial sequences provide valuable data regarding xenobiotics of the host species. For example, this is the first indication that *M. cinxia* butterflies are infected with microsporidia, an intracellular parasite that affects insect population dynamics (Kohler & Wiley 1992; Kohler & Hoiland 2001). Microarray analysis revealed marked differences among individuals in the transcript levels of microsporidia genes (Fig. 5), and we can now explore unanticipated issues such as the relationship between expression of specific host and parasite genes. This highlights the discovery and hypothesis-expanding aspects of a broad and deep sequencing approach, which in this case has the potential to provide insight regarding physiological mechanisms that may have important ecological consequences.
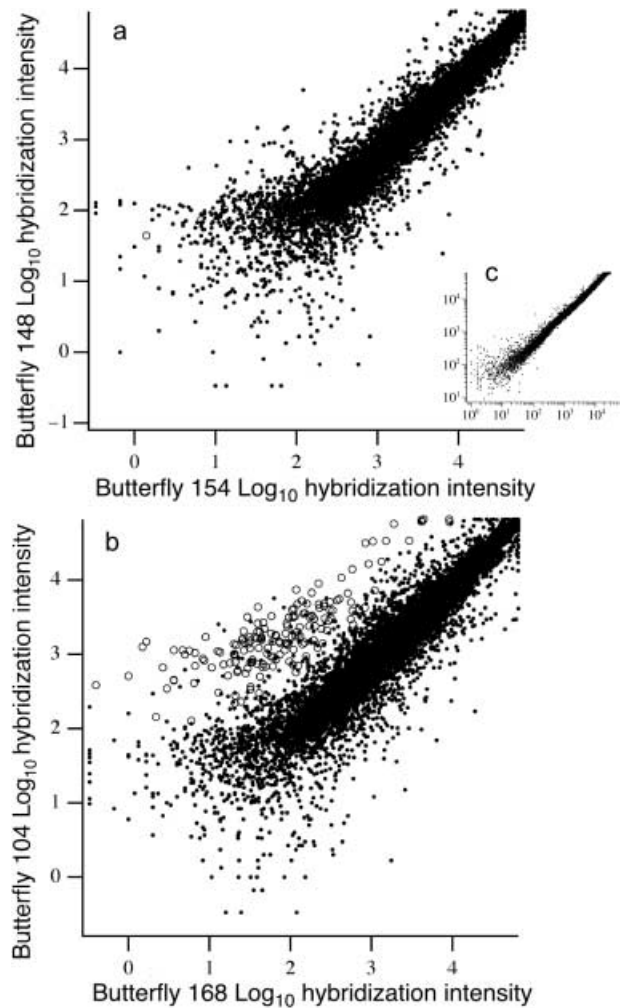

Fig. 5 Examples of 44K-feature Agilent arrays with probes made from 454 sequence data. (a,b) Comparison of raw hybridization intensity from the two channels of microarrays hybridized with indirectly labelled aRNA from whole abdomens of 2-day-old female *Melitaea cinxia* butterflies. Note the substantial increase in transcript level of microsporidia genes (larger open circles) in butterfly 104 compared to butterfly 168 (b), and the apparent absence of microsporidia in the two other individuals (a). (c) Repeatability of signal intensity (mean of three replicates per feature) across two microarrays hybridized with labelled abdominal aRNA from a single individual.

Discussion

We used samples from whole bodies to characterize the transcriptome of a nonmodel species using 454 pyro-sequencing and *de novo* assembly, and used those data to discover large numbers of SNPs and to construct a richly featured microarray. Although a precise estimate of transcriptome coverage is unattainable without full genomic sequence, we appear to have recovered over half, and potentially significantly more, of the *Melitaea cinxia* transcriptome. Xenobiotic sequences were also detected

and those data will be useful for developing studies on the interaction between butterflies and their microsporida parasites in natural populations. The assembled sequence data and singletons provided a rich source of material for construction of an oligonucleotide microarray that showed both high repeatability and ability to detect biological differences among individuals.

These successes need to be considered alongside limitations of the approach in comparison to established Sanger EST library sequencing methods. Had we not generated libraries from which we collected Sanger sequences for comparison purposes, we would not have clones. Without clones, one cannot readily obtain full-length sequence data for a given gene of interest, but must instead use the fragment data as the basis for a RACE protocol. Second, 454 sequences are located fairly evenly across the cDNA of a given gene (Weber *et al*. 2007), rather than being from one terminal end. Sequences that are distributed away from untranslated end regions are beneficial for obtaining BLAST alignments, but this scattered coverage also results in multiple fragments per gene, requiring further assessment in downstream analyses to assess their relationships. Finally, because 454 sequences are derived from both strands, the resulting sequence data have an unknown directional orientation. Directionality can be inferred from BLAST annotation, or as we did for unannotated genes, from the relative signal intensity of complementary microarray probes. Investigators choosing between approaches based solely on unoriented cDNA fragments (Hudson 2007) vs. clones should carefully consider the potential limitations imposed by these factors.

A challenge for any EST project is obtaining sufficient coverage of less abundant transcripts. We used normalized cDNA to reduce oversampling of abundant transcripts, but there remained substantial variation in average coverage depth. The average contig was fairly short (~200 bp), and even though 88% of the sequence reads assembled into contigs, there remained *c*. 60 000 singletons. Thus, even from a normalized cDNA pool, coverage of rarer transcripts was often thin. Singletons provided the only sequence that aligned to 1527 of the 6289 unique *Bombyx mori* orthologues identified with BLAST analysis, and because they had no redundant coverage, they contained a higher error and gap rate than the contigs. A recent upgrade of the 454 sequencer now provides reads averaging *c*. 230 bp and more reads per run; this suggests that coverage and assembly performance will increase, perhaps dramatically, above what we have described here. Lack of annotation for many of the genes discovered will remain a problem when working with nonmodel species, regardless of methodological approach or assembly quality.

Previous attempts at *de novo* transcriptome assembly from 454 sequences used different methods and achieved significantly lower average coverage (0.23× coverage vs.

our estimated 2.3×, Table S8; Cheung *et al*. 2006). High coverage is critical for joining of overlapping ESTs into larger contigs, but even low-covered genes can provide valuable data if full-length gene characterization is not the primary goal. Our annotated singletons (~80–100 bp after quality trimming) provided sufficient sequence length to design microarray probes, most of which produced detectable hybridization signals.

A key feature of 454 sequencing is that, compared to a similar dollar expenditure on Sanger sequencing, it yields redundant coverage for more genes, thereby allowing much more extensive SNP discovery. Until now, 454 transcriptome-wide SNP detection (as opposed to ultradeep coverage of single genes or amplicons) has been accomplished only by comparison with genome data (Bainbridge *et al*. 2006; Barbazuk *et al*. 2007). Our initial scan of gap-free BLAST-aligned regions in long and deeply covered contigs revealed a rich collection of SNPs, including 149 first and second codon position polymorphisms that are likely to change the amino acid sequence. We are presently constructing an algorithm that accomplishes assembly-wide discovery of high-quality SNPs using specified criteria (depth, quality score, avoidance of homopolymer runs, frequency, inferred codon position within annotated contigs) in *de novo* assemblies of 454 ESTs; we will report on this in a subsequent publication. The ability to accomplish transcriptome-wide SNP discovery using 454 sequence data is an important development because gene-associated SNPs are difficult and expensive to obtain by other methods but provide valuable data and markers for population biology, functional genomics, and conservation (Rosenberg & Nordborg 2002; Balding 2006; Kohn *et al*. 2006), including nonmodel species in which SNP proximity and chromosomal location may be inferred by assuming synteny with known genomic sequences of related model taxa (Joron *et al*. 2006; Yasukochi *et al*. 2006).

The Glanville fritillary butterfly has been studied extensively from an ecology and population biology perspective (Hanski 1999; Ehrlich & Hanski 2004). Like most free-living species, it has lacked genetic and genomic resources necessary for mechanistic study. Here, we have demonstrated that it is possible to use 454 pyrosequencing to rapidly characterize a draft transcriptome and use it to create a microarray for large-scale functional genomics. These tools will allow us to determine how variation in gene sequences and gene expression are associated with key ecological features of this species, including flight ability, dispersal, fecundity, and the impact of metapopulation parameters on these traits. The methods described here can be used for potentially any species, thereby narrowing the gap between approaches based on model organisms with rich genetic resources vs. species that are most tractable for ecological and evolutionary studies (i.e. the Krogh principle; Krebs 1975; Wayne & Staves 1996).

Acknowledgements

References

Adams MD *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.

Ashburner M *et al.* (2000) Gene ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25–29.

Auboeuf D, Batsche E, Dutertre M, Muchardt C, O'Malley BW (2007) Coregulators: transducing signal from transcription to alternative splicing. *Trends in Endocrinology and Metabolism*, **18**, 122–129.

Bainbridge MN *et al.* (2006) Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics*, **7**, 246.

Balding DJ (2006) A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, **7**, 781–791.

Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS (2007) SNP discovery via 454 transcriptome sequencing. *The Plant Journal*, **51**, 910–918.

Beldade P, Rudd S, Gruber JD, Long AD (2006) A wing expressed sequence tag resource for *Bicyclus anynana* butterflies, an evo-devo model. *BMC Genomics*, **7**, 130.

Boeckmann B *et al.* (2003) The SWISS-PROT protein knowledgebase and its (Suppl.) TrEMBL in 2003. *Nucleic Acids Research*, **31**, 365–370.

Bouck A, Vision T (2007) The molecular ecologist's guide to expressed sequence tags. *Molecular Ecology*, **16**, 907–924.

Cheung F *et al.* (2006) Sequencing *Medicago truncatula* expressed sequenced tags using 454 Life Sciences technology. *BMC Genomics*, **7**, 272.

Ehrlich PR, Hanski I (2004) *On the Wings of Checkerspots: A Model System For Population Biology*. Oxford University Press, Oxford, UK.

Emrich SJ, Barbazuk WB, Li L, Schnable PS (2007) Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Research*, **17**, 69–73.

Goldberg SMD *et al.* (2006) A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proceedings of the National Academy of Sciences, USA*, **103**, 11240–11245.

Haag CR, Saastamoinen M, Marden JH, Hanski I (2005) A candidate locus for variation in dispersal rate in a butterfly metapopulation. *Proceedings of the Royal Society B: Biological Sciences*, **272**, 2449–2456.

Hanski I (1999) Habitat connectivity, habitat continuity, and metapopulations in dynamic landscapes. *Oikos*, **87**, 209–219.

Hanski I, Saccheri I (2006) Molecular-level variation affects population growth in a butterfly metapopulation. *PLoS Biology*, **4**, e129.

Hanski I, Eralahti C, Kankare M, Ovaskainen O, Siren H (2004) Variation in migration propensity among individuals maintained by landscape structure. *Ecology Letters*, **7**, 958–966.

Hanski I, Saastamoinen M, Ovaskainen O (2006) Dispersal-related life-history trade-offs in a butterfly metapopulation. *Journal of Animal Ecology*, **75**, 91–100.

Hudson ME (2007) Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Notes*, doi: 10.1111/j.1471-8286.2007.02019.x

Hughes TR *et al.* (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nature Biotechnology*, **19**, 342–347.

Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM (2007) Accuracy and quality of massively-parallel DNA pyrosequencing. *Genome Biology*, **8**, R143.

Joron M *et al.* (2006) A conserved supergene locus controls colour pattern diversity in *Heliconius* butterflies. *PLoS Biology*, **4**, 1831–1840.

Kersey P, Hermjakob H, Apweiler R (2000) VARSPLIC: alternatively-spliced protein sequences derived from SWISS-PROT and TrEMBL. *Bioinformatics*, **16**, 1048–1049.

Kim E, Magen A, Ast G (2007) Different levels of alternative splicing among eukaryotes. *Nucleic Acids Research*, **35**, 125–131.

Kohler SL, Hoiland WK (2001) Population regulation in an aquatic insect: the role of disease. *Ecology*, **82**, 2294–2305.

Kohler SL, Wiley MJ (1992) Parasite-induced collapse of populations of a dominant grazer in Michigan streams. *Oikos*, **65**, 443–449.

Kohn MH, Murphy WJ, Ostrander EA, Wayne RK (2006) Genomics and conservation genetics. *Trends in Ecology & Evolution*, **21**, 629–637.

Krebs HA (1975) The August Krogh principle for many problems there is an animal on which it can be most conveniently studied. *Journal of Experimental Zoology*, **194**, 221–226.

Marden JH (2006) Quantitative and evolutionary biology of alternative splicing: how changing the mix of alternative transcripts affects phenotypic plasticity and reaction norms. *Heredity*, 2006 September 27; [Epub ahead of print].

Margulies M *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

Mita K *et al.* (2003) The construction of an EST database for *Bombyx mori* and its application. *Proceedings of the National Academy of Sciences, USA*, **100**, 14121–14126.

Moore MJ *et al.* (2006) Rapid and accurate pyrosequencing of angiosperm plastid genomes. *BMC Plant Biology*, **6**, 17.

Nagaraj SH, Gasser RB, Ranganathan S (2007) A hitchhiker's guide to expressed sequence tag (EST) analysis. *Briefings in Bioinformatics*, **8**, 6–21.

Orsini L, Pajunen M, Hanski I, Savilahti H (2007) SNP discovery by mismatch-targeting of Mu transposition. *Nucleic Acids Research*, **35** (6), e44.

Pajares MJ *et al.* (2007) Alternative splicing: an emerging topic in molecular and clinical oncology. *Lancet Oncology*, **8**, 349–357.

Papanicolaou A, Joron M, McMillan WO, Blaxter ML, Jiggins CD (2005) Genomic tools and cDNA derived markers for butterflies. *Molecular Ecology*, **14**, 2883–2897.

Paschall JE *et al.* (2004) FunnyBase: a systems level functional annotation of *Fundulus* ESTs for the analysis of gene expression. *BMC Genomics*, **5**, 96.

Poinar HN, Schwarz C, Qi J *et al.* (2006) Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science*, **311**, 392–394.

Rosenberg NA, Nordborg M (2002) Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics*, **3**, 380–390.

Rudd S (2003) Expressed sequence tags: alternative or complement to whole genome sequences? *Trends in Plant Science*, **8**, 321–329.

Saastamoinen M (2007a) Life-history, genotypic, and environmental

correlates of clutch size in the Glanville fritillary butterfly. *Ecological Entomology*, **32**, 235–242.

Saastamoinen M (2007b) Heritability of dispersal rate and other life history traits in the Glanville fritillary butterfly. *Heredity*. September 5; [Epub ahead of print].

Shagin DA *et al.* (2002) A novel method for SNP detection using a new duplex-specific nuclease from crab hepatopancreas. *Genome Research*, **12**, 1935–1942.

Toth AL, Varala K, Newman TC *et al.* (2007) Wasp gene expression supports an evolutionary link between maternal behavior and eusociality. *Science*, September 27; [Epub ahead of print].

Trombetti GA, Bonnal RJP, Rizzi E, De Bellis G, Milanesi L (2007) Data handling strategies for high throughput pyrosequencers. *BMC Bioinformatics*, **8**, S22.

Wang J, Xia Q, He X *et al.* (2005) SilkDB: a knowledgebase for silkworm biology and genomics. *Nucleic Acids Research*, **33**, D399–D402.

Wayne R, Staves MP (1996) The August Krogh principle applies to plants. *Bioscience*, **46**, 365–369.

Weber APM, Weber KL, Carr K, Wilkerson C, Ohlrogge JB (2007) Sampling the *Arabidopsis* transcriptome with massively parallel pyrosequencing. *Plant Physiology*, **144**, 32–42.

Whitfield CW *et al.* (2002) Annotated expressed sequence tags and cDNA microarrays for studies of brain and behavior in the honey bee. *Genome Research*, **12**, 555–566.

Wicker T *et al.* (2006) 454 sequencing put to the test using the complex genome of barley. *BMC Genomics*, **7**, 275.

Wondji CS, Hemingway J, Ranson H (2007) Identification and analysis of single nucleotide polymorphisms (SNPs) in the mosquito *Anopheles funestus*, malaria vector. *BMC Genomics*, **8**, 5.

Xia QY *et al.* (2004) A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science*, **306**, 1937–1940.

Yasukochi Y, Ashakumary LA, Baba K, Yoshido A, Sahara K (2006) A second-generation integrated map of the silkworm reveals synteny and conserved gene order between lepidopteran insects. *Genetics*, **173**, 1319–1328.

Zhu YY, Machleder EM, Chenchik A, Li R, Siebert PD (2001) Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *BioTechniques*, **30**, 892–897.

Zhulidov PA *et al.* (2004) Simple cDNA normalization using kamchatka crab duplex-specific nuclease. *Nucleic Acids Research*, **32**, e37.

Zhulidov PA *et al.* (2005) A method for the preparation of normalized cDNA libraries enriched with full-length sequences. *Russian Journal of Bioorganic Chemistry*, **31**, 170–177.

Cris Vera is a graduate student applying bioinformatics tools to next generation sequencing projects and functional genomics. Chris Wheat is a post-doc who integrates molecular evolution and functional genomics with physiology and ecology. Howard Fescemyer studies genetic mechanisms of insect development, endocrinology and reproductive physiology. Mikko Frilander studies post-transcriptional gene regulation and alternative splicing in eukaryotes. Doug Crawford studies variation in gene expression and its physiological and evolutionary importance. Ilkka Hanski studies population regulation, cyclic population dynamics, mechanisms of coexistence in communities, and metapopulation biology. Jim Marden studies mechanisms that create variability in organismal performance, and the ecological and evolutionary consequences thereof.

Supplementary material

The following supplementary material is available for this article:

Table S1 Number of reads and nucleotides produced by two 454 sequencing runs

Table S2 Sanger EST statistics

Table S3 Summary statistics for BLAST of assembled *Melitaea cinxia* Sanger sequences against *M. cinxia* 454 contigs and singletons. All BLAST results refer to hits with bitscores greater than or equal to 45. Percentage length coverage refers to the total extent of Sanger sequences covered by aligned 454 sequences. Per cent identity is reduced by polymorphism and reduced accuracy of Sanger sequencing at end regions; it is not a valid measure of the 454 error rate. Alignment lengths refer to nucleotides (not deduced amino acids as in Fig. 3 and Table 2)

Table S4 Summary statistics for BLAST analyses of *Melitaea cinxia* 454 contigs and singletons against predicted *Bombyx mori* proteins, Uniprot proteins, and translated clustered ESTs in Butterflybase. A minimum bitscore of 45 was used as the cutoff threshold for all of these analyses. Sequences that had best hits in Uniprot that were non-metazoans (*N* = 716) are excluded. Values in parentheses are medians

Table S5 GO assignments using only Arthropod BLASTX alignments

Table S6 Taxonomic group of the BLAST subject species for 454 contigs and singletons (*N* = 699) that had significant (bitscore > 45) best BLAST alignments to apparent xenobiotic genes

Table S7 Settings used in SEQMAN PRO 7.1 to assemble 454 ESTs

Table S8 Transcriptome space and coverage comparison among relevant taxa. Expected coverage refers to the average coverage depth predicted for that size genome from our size 454 sample, or in the case of *Medicago truncatula*, to the average coverage depth obtained in that study (Cheung *et al.* 2006)

Fig. S1 Gel image of two cDNA collections (A is larval/pupal material; B is adult) made from poly(A +) mRNA, before and after cDNA normalization. Size markers are shown in the lane at the left; normalized cDNA lanes are labelled N-cDNA. Note the high relative abundance of about 10 eretranscripts (distinct bands) in the cDNA lanes and complete absence thereof after normalization.

Fig. S2 Examples of BLAST alignments of *Melitaea cinxia* assembled Sanger ESTs against *M. cinxia* 454 contigs. (a) An alignment that is representative of the average alignment length and per cent identity (see Table S3). Note the variable presence of what appears to be an alternatively spliced microexon (single codon). (b) The longest Sanger vs. 454 alignment.

Fig. S3 Upper, schematic showing how local clusters were often not joined together due primarily to lack of critical sequence reads necessary for joining clusters into one full length, or nearly full-length contig. Naturally occurring variation in the form of polymorphism and alternative splicing, and bias towards increased coverage of transcript end regions (Bainbridge *et al.* 2006; see Fig. S4a) can hamper assembly, even where there is high coverage and

low rate of sequencing errors. Lower, within clusters of multiple reads, polymorphic sites, or SNPs, are readily identified. The right side of this alignment shows an example of how sequencing error is ignored in the consensus sequence (underlined sequence above the alignment).

Fig. S4 (a) Example of sampling biases (Bainbridge *et al*. 2006) present in 454 sequencing. Sequence ends, particularly the 5′ end, tend to be over-represented in 454 ESTs, with under-representation of the immediate flanking region. This pattern is discernable in the most heavily covered contigs in which there was enough sampling of the end-flanking region to join the end to the remainder of the contig. (b) Relatively even coverage of the remainder of the same contig shown in *A*. (c) Bimodal distribution of coverage depth as a function of contig length. The upward sloping pattern is expected (i.e. contigs become longer when there is more sequence coverage, up to the point of full length assembly). The circled portion of the plot contains contigs in which there was high coverage but little or no length extension. Few of these generated BLAST hits and thus appear to be over-sampled end regions [presumably primarily UTR (untranslated region)] that lacked sufficient flanking coverage to allow them to assemble with the remainder of ESTs for that gene.

Fig. S5 (a–c) Frequency distributions of $\log_{10}$-transformed median spot intensity (corrected for local background and with a constant added to make all values positive for log transformation) of microarray probes. $N = 602$ negative control probes; $N = 125\,655$ probes for contigs and singletons that had significant BLAST hits; $N = 113\,862$ unannotated contigs. This initial test array had low

signal intensity overall [note lack of saturated spots (intensity values of 4.8)] and thus yields conservative estimates of the number of quality probes. The greater number of very low signals in B, C compared to A is due to the extreme difference in sample size. (d) Distribution of $\log_{10}$-transformed median spot intensity for the lesser (upper plot) and greater (lower plot) performing probe orientation for all probes of unannotated genes for which one probe had an intensity of at least 2.5 (i.e. a strong signal in at least one probe orientation). Signals of the lesser performing probes cluster at just above the intensities obtained from negative control probes, with an average 10-fold difference among these differently orientated probe pairs.

Fig. S6 Probes from 454 sequence generally produce comparable microarray spot intensity as probes designed from Sanger sequence of the same gene ($r^2 = 0.57$; $N = 892$). Diagonal is the line of identity. There was no systematic tendency for either type of probe to yield better performance.

Appendix S1 Materials and methods

This material is available as part of the online article from:
http://www.blackwell-synergy.com/doi/abs/
10.1111/j.1365-294X.2007.03666.x
(This link will take you to the article abstract).

Please note: Blackwell Publishing are not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.