

Rapidly developing functional genomics in ecological model systems via 454 transcriptome sequencing

Christopher W. Wheat

Received: 20 April 2008 / Accepted: 22 September 2008 / Published online: 18 October 2008
© Springer Science+Business Media B.V. 2008

Abstract Next generation sequencing technology affords new opportunities in ecological genetics. This paper addresses how an ecological genetics research program focused on a phenotype of interest can quickly move from no genetic resources to having various functional genomic tools. 454 sequencing and its error rates are discussed, followed by a review of de novo transcriptome assemblies focused on the first successful de novo assembly which happens to be in an ecological model system (the Glanville fritillary butterfly). The potential future developments in 454 sequencing are also covered. Particular attention is paid to the difficulties ecological geneticists are likely to encounter through reviewing relevant studies in both model and non-model systems. Various post-sequencing issues and applications of 454 generated data are presented (e.g. database management, microarray construction, molecular marker and candidate gene development). How to use species with genomic resources to inform study of those without is also discussed. In closing, some of the drawbacks of 454 sequencing are presented along with future prospects of this technology.

Keywords 454 Transcriptome sequencing · Ecological genetics · Functional genomics · Glanville fritillary butterfly · EST

Introduction

Ecological genetics has long sought to identify and mechanistically understand the role of specific genes in ecology and evolution (Ellegren and Sheldon 2008; Endler 1986; Gillespie 1991; Lewontin 1974). Although the genomics revolution has greatly facilitated this process, especially in genomic model systems and their close relatives (Feder and Mitchell-Olds 2003; Mitchell-Olds et al. 2007), the number of model ecological systems which have been able to gain such insights has been limited (e.g. Abzhanov et al. 2006; Nachman et al. 2003), due to a combination of high costs, small research communities, and a need for truly integrated scientific research programs. Recent technological advances in high throughput sequencing have greatly lowered the hurdles for genomic tool development which facilitate functional genomic insights (Ellegren 2008; Margulies et al. 2005).

Here the performance and implications of next generation sequencing advances are discussed. Emphasis is placed on the 454 sequencing technology, covering recent advances in model genomic systems and representative model ecological systems with previously limited genomic resources. This paper addresses how an ecological research program focused on a phenotype of interest can quickly move from no genetic resources to developing mechanistic understanding, using the Glanville fritillary butterfly as a proof of concept example (Ellegren 2008; Vera et al. 2007). This paper is written for the ecologist who knows what a microsatellite or AFLP is, but not cDNA or a contig, and is a good companion paper to the recent review on traditional EST libraries and their uses (Bouck and Vision 2007). Concepts and findings are fully discussed using the relevant terminology with their common abbreviations, which I have tried to define at a general level.

C. W. Wheat (✉)
Metapopulation Research Group, Department of Biological and Environmental Sciences, University of Helsinki,
P.O. Box 65, Viikinkaari 1, Helsinki 00014, Finland
e-mail: chris.wheat@stanfordalumni.org

C. W. Wheat
Department of Biology, Pennsylvania State University,
University Park, PA 16802, USA

This paper is organized into 11 sections, followed by a conclusion. Beginning with a focus on the importance of research questions and how these determine the tools that need to be developed (*Research questions and tools*), 454 sequencing of the transcriptome (*Transcriptome sequencing*) is then discussed followed by a detailed discussion of its error rates compared to Sanger sequencing (*454 sequencing error rates*). A review of previous attempts at de novo transcriptome assembly are presented, focusing on a successful assembly in an ecological model system (*Transcriptome assembly*). Particular attention is paid to the difficulties the ecological geneticist is likely to face, through both a detailed review of relevant studies in model and non-model systems when considering how to assess the transcriptome coverage of a given 454 sequencing run. Various post-sequencing applications of data are then presented, such as annotation and assessing how much of the transcriptome was sequenced (*Annotation & assessing transcriptome coverage*), microarray construction (*Microarrays*), molecular markers (*Molecular markers: SNPs, microsatellites, and EPICs*), candidate gene development (*Candidate genes*), and genomic scans (*Genomic scans*), all with a focus on an ecological model system. A discussion of the importance of understanding the molecular resources available for closely related species then highlights how these can be used for the focal species without such resources (*Genomic referencing summary*). In closing, some of the drawbacks of 454 sequencing (*Drawbacks of 454 sequencing*) are presented followed by a conclusion.

Research questions and tools

Ecological genetics research asks what are the genes, the performances of the resulting molecular and organismal phenotypes, and the selective regimes that result in differential reproductive success in the wild (Feder and Mitchell-Olds 2003; Feder and Watt 1992). Organismal phenotypes may be morphological, behavioral, physiological, or combinations of these. For example, trying to understand the genetic basis of variation in dispersal among individual butterflies of a given species could encompass variation in wing size, a decision to disperse, and the energy to sustain flight. Trying to predict the genetic architecture underlying a complex phenotype such as differential dispersal ability is thus difficult even with functional genomic study in a relevant model system (e.g. *Drosophila*). While regulatory variation may have a larger role in some aspects of phenotypic evolution compared to coding (Carroll 2005; Wray 2007), the current limited understanding of adaptive trait architecture warrants incorporating the paucity of information on this topic into research project design (Ellegren and Sheldon 2008). Thus,

a more focused research question may ask what are the genes whose expression affects the phenotype of interest, or which genes harbor coding variation with relevant phenotypic consequences. To address these questions in ecological model systems, molecular tools usually need to be developed where none previously existed.

In a dream with unlimited time and funds, a researcher might be tempted to obtain the whole genome sequence (WGS) of their study organism. Upon waking, having such data poses its own problems in terms of assembly, annotation, and navigation through the vast amount of accumulated data. One needs a focused research community that is going to invest into such a genomic project and given the scattered nature of ecological model systems, few research communities are large enough for such a task (Bouck and Vision 2007). More importantly, what could one even get from a sequenced genome? Since genome sequencing ideally uses DNA from a single individual or inbred strain, little to no genetic variation information is recovered. As a result, one is able to assess genomic architecture (e.g. copy number of genes, their relative locations, intron size variation, codon bias, simple repeat regions, etc.), but not variation among individuals. Gene sequences could be predicted from WGS data, allowing the design of microarrays and primers for genomic scans of genetic variation via re-sequencing. But, given the low percentage of coding genes in most genomes (e.g. <2% in humans), this is a poor return on time and money investment.

What are the molecular tools then that will aide ecological research groups? Most ecological genetics labs began using microsatellites to understand population structure, which has now giving way to single nucleotide polymorphism (SNP) variation as a marker of choice (due to better understood mutation models, genomic distribution, and availability; (Morin et al. 2004)). Both of these markers can also be used for detecting loci under selection via genomic scans, QTL, and association studies (Luikart et al. 2003; Morin et al. 2004; Storz 2005; Vos et al. 1995). Such studies are not biased by assuming a priori a specific architecture underlying a given phenotype (e.g. such as assuming a larger role of expression or coding variation), because they blindly query the genome to find chromosomal regions associated with the studied phenotype (Slate 2005). However, quickly attaining large numbers of microsatellites or SNP loci can be difficult and expensive, especially in certain taxa (e.g. microsatellite difficulties in Lepidoptera (Van't Hof et al. 2007; Zhang 2004)).

Amplified fragment length polymorphisms (AFLPs) are a potential solution to these issues, quickly providing numerous polymorphic markers scattered across the genome (Vos et al. 1995). However, AFLP markers are essentially species or even population specific, as they can't be integrated with findings from other taxa in terms of

polymorphism, chromosomal location, etc, and as dominant biallelic markers contain less information than co-dominant biallelic such as SNPs (Morin et al. 2004). In sum, developing molecular markers which can use the WGS information from a genomic reference species (GRS) is important (i.e. a better investment).

A GRS is the evolutionarily closest or most relevant species to the focal species for which there exist genomic resources, such as many sequenced genes or WGS. Having a good GRS allows microsatellites or SNPs developed in the focal species to be potentially more useful than AFLPs (Bouck and Vision 2007). For example, knowing you have SNPs within specific genes, which have orthologs in the GRS, can allow genomic insights to be quickly integrated across taxa due to shared gene order, or synteny. Such a technique was used to develop markers for mapping in model systems such as mammals and flowering plants (Bouck and Vision 2007; Fulton et al. 2002; Lyons et al. 1997). Recently, this technique was used to exploit growing insight into the genetics of wing pattern variation in the Müllerian mimicry model system of *Heliconius* butterflies, finding that a similar chromosomal region has large and different effects on wing patterns across species (Joron et al. 2006).

Returning back to genomic tool development, quickly gaining access to the ~3% of an organism's genome that actually codes for genes provides the necessary data for developing the tools discussed above, and much more (Bouck and Vision 2007). This can be accomplished by transcriptome sequencing, which provides direct access to the mRNA sequence containing coding gene sequence as well as both the 5' and 3' flanking untranslated regions (UTR) (the 3' end of the UTR is a long stretch of A's called the poly A tail). Sequence from mRNA can be used for quantifying global gene expression (via designed microarrays) or genome wide coding variation (via sequencing 100's of genes (Bouck and Vision 2007; Schmid et al. 2005a). When the mRNA material is a pool of outbred individuals, transcriptome sequencing can also provide data for finding 100's of microsatellites and 1000's of SNPs, which can be located in either coding or the more variable UTR regions (e.g. Beldade et al. 2007; Bouck and Vision 2007; Kantety et al. 2002; Picoult-Newberg et al. 1999). Recent technological advances have brought transcriptome sequencing, which was traditionally labor intensive and costly, to within reach of any research group during a normal grant period (3 years) and funding range (e.g. Vera et al. 2007).

Transcriptome sequencing

Traditionally, the mRNA transcript pool (all the expressed genes isolated from a given tissue) has been sequenced by constructing a cDNA library (Bouck and Vision 2007).

Genes are transcribed (expressed), which makes mRNA, and the mRNA is isolated from high quality tissue. The DNA complement of the mRNA strand, called cDNA, is made using reverse transcriptase which reads the mRNA 3'–5'. After removal of the mRNA, this single stranded cDNA pool is made double stranded and then cloned into plasmids, which are transformed into bacteria. A dilution of this bacteria can then be plated such that individual bacterial cells (with one plasmid each) form unique colonies. Each individual bacterial colony thus contains a uniquely cloned mRNA strand (a cDNA clone). Individual cDNA clones are selected, grown and plasmids isolated or directly PCR'd using primers on the plasmid. These are then sequenced, generally from the 5' end as this avoids the poly A tail at the 3' end, providing DNA sequence of the 5' UTR and coding region of a gene, as well as the 3' UTR if the mRNA is short enough. This collection of bacteria, plasmids, or DNA data can be called an expressed sequence tag (EST) library, but in this paper EST refers to a given sequenced cDNA, rather than the physical library itself.

If the goal is maximal transcriptome coverage instead of transcripts specific to a given tissue type, then the mRNA from many different tissue types across sexes and developmental stages need to be isolated and pooled. Additionally, many genes are only induced upon environmental stimuli which, if relevant, would need to be incorporated into tissue harvesting protocols. Also note that the reverse transcription reaction used to make the cDNA does not always proceed to the end of the mRNA strand, due to the formation of secondary structures by the RNA which can block the progress of the reverse transcriptase. Thus many cDNA clones will be only partial mRNA fragments (but there are ways around this problem such as using higher temperatures for the reverse transcription reaction and SMART cDNA synthesis; (Zhu et al. 2001)). Coupled with the read lengths of Sanger sequencing (~700 bp), the net result is that many EST sequences, again primarily from the 5' end of the cDNA, will likely only capture some portion of the coding gene as many genes are much longer than average Sanger sequencing length.

Cloning of the mRNA pool results in three potential biases in cDNA libraries. First, not all genes are expressed at equal levels, with, for example, housekeeping genes such as ribosomal proteins being highly expressed compared to the vast majority of other genes (e.g. the mRNA of 20 housekeeping genes alone can constitute up to 50% of the mRNA isolated from a given tissue). Ongoing technological development offers several different techniques for the “normalization” of such mRNA pools and requires molecular biology experience. Normalization is however commercially available (e.g. Google search “cDNA library

normalization”). Normalization involves taking a mRNA pool having few genes of many transcripts, and many genes of few transcripts, and modifying it such that all genes have a roughly similar number of transcripts (e.g. reducing a 10,000× difference in mRNA across genes to a ~10× difference; Bonaldo et al. 1996; Zhulidov et al. 2004). When sequencing from a normalized transcript pool, each successive randomly sequenced cDNA has a greater probability of being a new gene, rather than a previously sequenced transcript. Thus, sequencing a normalized cDNA library is a better return on investment when the goal is transcriptome coverage. But keep in mind that normalization does not result in all genes having an equal probability of being sequenced. Rather, normalization creates a new bias in transcript recovery probabilities such that they are more evenly distributed across genes.

The second bias of cDNA libraries is the potential of the mRNA transcript in plasmids to be partially expressed in their bacterial cells with lethal effects. These cells would then not proliferate and these transcripts would be lost to future sequencing (e.g. Weber et al. 2007). This problem has largely been overcome through use of vectors that don't express their inserts, although the bias still exists in old cDNA library databases. Finally, smaller cDNA fragments are over represented compared to larger ones, due to both the difficulty of obtaining full length cDNA from reverse transcription of long genes and the higher transformation efficiency of smaller plasmids. Size selection of cDNA is generally performed in an attempt to gain access to longer, full length transcripts, with some researchers cloning different size groups independently (e.g. Lévesque et al. 2003).

Making and sequencing cDNA libraries can be expensive. cDNA libraries are almost always Sanger sequenced using purified plasmids from the 5' end (which avoids the poly A tail at the 3' end, but for an example of 3' sequencing see Beldade et al. 2007), with most labs keeping the sequenced plasmid, or bacterial line, as frozen stock for later eventual full length or reverse sequencing. Costs can thus include plasmid purification, sequencing, and storage of plasmids and/or bacterial stocks, all of which include substantial handling as sequencing reads accumulate. cDNA libraries are usually sequenced with 1000's of reads, ranging from 5000 to >100,000 reads. Given the quality of normalization and tissue used, 10,000 such reads may find as few as 1000 and potentially >5000 unique genes. Thus, with sequencing costs about \$4 per run and an optimal cDNA library, getting at least partial coverage of 5000 unique genes would cost \$40,000, plus potential additional costs mentioned above.

In 2005 next generation sequencing became a reality, with massively parallel pyrosequencing. This new technique was able to take a non-cloned pool of DNA and

within 4 h sequence ~300,000 reads of ~110 bp in length each, producing a total of ~33 × 10⁶ bp or 33 Mb (Margulies et al. 2005). Using non-cloned DNA is an important advancement, in addition to the increase throughput and speed, since this avoids the potential toxicity of inserts and fragment size biases of traditional cDNA libraries discussed previously. The techniques involved in 454 sequencing have been nicely presented elsewhere. They involve a fundamental shift away from electrophoretic separation of dye terminated DNA fragments to a method of massively parallel record keeping of iterative nucleotide extension called flow cell sequencing (Holt and Jones 2008; Hudson 2008; Mardis 2008; Margulies et al. 2005). Quite simply, the process begins with a DNA pool that is randomly sheared followed by massively parallel sequencing of the resulting individual DNA fragments, utilizing advances in microscale DNA amplification and detection of differential nucleotide incorporation (Margulies et al. 2005).

In order to appreciate these advances, consider an idealized, normalized, average transcriptome of 20,000 genes with an average gene length of 1500 bp or 1.5 kb. This contains about 30 Mb of unique DNA. For illustration purposes, a back of the envelope calculation predicts a single 454 run could provide ~1 × transcriptome coverage of this pool (i.e. 33/30 = 1.1). A single run using this technology currently costs ~\$10,000, providing a sharp contrast to the traditional EST sequencing outlined above. This cost estimate was based on using a university core facility and costs going directly through 454/Roche are currently ~\$20,000, which is still a significant reduction in cost compared to Sanger sequencing. Perhaps a more impressive show of this technology was the recent sequencing of James Watson's entire genome, which took only 2 months to sequence the 6 gigabases of his DNA to 7.4× coverage, all at 1/100 the costs and a fraction of the time of traditional capillary sequencing (Wheeler et al. 2008).

Other high throughput next generation sequencing technologies are available, such as SOLiD (Applied Biosystems Inc.) and Solexa (Illumina Solexa Inc.), which can generate a little over 2 million ~30 bp fragments per run, which is roughly a third of the human genome (for comparisons across these and other methods see these reviews (Holt and Jones 2008; Hudson 2008; Mardis 2008)). While providing massively more data, the performance for de novo transcriptome assembly using such short read lengths has not yet been assessed. However, these and other next generation sequencing technologies are increasing their throughput, read length, lowering costs, and improving performance. For example, by mid 2008, 454 sequencing technology (now called GS FLX) currently has an average run time of 8 hours, generating ~400,000 reads, each

~240–250 bp long, totalling >100 Mbp of data. By 2009, they appear likely attain from a 12 h run an average length of 450 bp for potentially near a million reads. Again, using a back of the envelope calculations, this final advance would provide ~15× coverage of the idealized transcriptome presented above from a single GS FLX run.

A final point to add is that several methods (e.g. Solexa and 454) now offer paired end sequencing (Holt and Jones 2008). This technique keeps track of the sequencing results from opposite ends of individual DNA strands and this helps in reconstructing the sequenced template (e.g. Korbel et al. 2007). 454 Life Science currently offers 100 bp reads from either end of a fragment up to 3,000 bp long, but this is likely to upgrade to about 175 bp reads on fragments up to 16 kb. Again, predicting which of the next generation sequencing techniques will perform best for a given sequencing project even a year from now is impossible as these and other technologies, and assembly methods, are rapidly advancing (Holt and Jones 2008).

454 sequencing error rates

But how does the error rate of this next generation sequencing technology compare to traditional Sanger EST sequencing? The next generation sequencing chemistry and base calling methods are profoundly different from traditional Sanger sequencing (Holt and Jones 2008; Margulies et al. 2005). Sanger sequencing reads individual bases directly, while the 454 sequencing method does not. Rather, the 454 method detects the incorporation of nucleotides in real time via pyrophosphate release during polymerase extension, which makes a flash of light. It does this for each of nearly 1.5×10^6 wells on a plate, each of which has a clonal pool of PCR amplified DNA.

The 454 method alternates the available nucleotides for incorporation across all wells and records an A from a pyrophosphate flash when A's are available, and then the process continues cycling through the different nucleotides and recording in which order they are incorporated (for each well). While incredibly fast and able to be miniaturized, the drawback of this method is that regions of DNA having several of the same nucleotides together, or a homopolymer run (e.g. AAA or AAAAA), result in essentially one nearly simultaneous pyrophosphate flash (the incorporation is very fast; (Hudson 2008)). Although the intensity of this flash does correlate with the number of nucleotides incorporated, this is where the typical errors for 454 occur with an over or under estimation of these run lengths or errors in regions next to them (Moore et al. 2006; Wicker et al. 2006). Thus, most publications on 454 sequencing report the majority of their errors associated with homopolymer runs of 5 or more (Margulies et al.

2005), with some reporting an additional AT bias of such runs (Moore et al. 2006; Wicker et al. 2006). Other sources of error come from mixed DNA pools in wells, misincorporation of nucleotides, etc. (see Holt and Jones 2008; Huse et al. 2007). In sum, the significant savings of time and money that 454 genome sequencing provides comes at a cost of a modest increase in error rate compared to traditional Sanger genome sequencing (~0.04% in 454 sequencing vs. 0.01% in Sanger sequencing; (Ewing and Green 1998; Margulies et al. 2005; Moore et al. 2006)). From these numbers alone, it appears that concerns about error rate differences are overblown, for what is the difference between 99.96% and 99.99% accuracy?

To discuss this, we need to consider the depth of coverage in these various studies and the statistics reported. Coverage refers to the number of sequences that cover a given region of sequence. From the perspective of a single bp within a stretch of DNA, this bp may have only one or two sequence reads (shallow or low coverage) or upwards of 5 or more (deep or high coverage) (Fig. 1). With coverage in mind, let us return to studies that have found very low error rates. Margulies et al. (2005) report a very low error rate, 0.04%, for their 454 sequencing and de novo assembly of 96% of a bacterial genome (580 kb). A similar low error rate, 0.043%, was reported from sequencing and de novo assembly of two plant plastid genomes (162 kb and 157 kb; Moore et al. 2006). Both studies were able to estimate their error rates by comparison back to Sanger sequence data, for some (Moore et al. 2006) or all of their sequenced genome (Margulies et al. 2005).

Average coverage for these genome projects was extremely high, ~40× in the bacterial genome and ~20× in the plastid example (i.e. each bp of the assembled genomes had at least 40 or 20 independent sequence reads per bp, respectively). Each of these independent sequence reads can be aligned with similar sequence to form a group of overlapping sequence, the consensus of which is called a contig. Having high coverage per bp allows for sequencing errors to be swamped by correct sequence during contig assembly, and thus, when contigs are used in the de novo genome assembly, error rates drop significantly and these are what are reported in the paragraph above.

In order to assess individual 454 sequence reads Margulies et al. (2005) compared their 454 sequence data at both the individual sequence and contig level with the known genome sequence of their bacteria. Individual sequence reads had an error rate of 3.3% for insertion and deletions and 0.5% for incorrect bp substitutions, with a summed error rate of ~400 bp in 10 kb. These errors increased dramatically at longer sequence lengths (i.e. >80 bp). This error rates drops significantly when the consensus sequences from contig assemblies are used instead, to 3 bp errors in 10 kb, with a noticeable decrease

type of coverage a given study requires depends upon the desired results and the diversity of the DNA within the sample. Comparison across several studies suggest that broad, accurate coverage of DNA material by current 454 sequencing performance can be achieved by $\sim 15\times$ coverage (Margulies et al. 2005; Vera et al. 2007; Wicker et al. 2006). As stated earlier, such a $15\times$ coverage of our idealized transcriptome may be achieved by a single 454 run sometime in 2009. On one hand, increased read lengths and accuracy may significantly decrease this required depth of coverage. However, a desire for SNPs will require substantial coverage. Considering transcriptome sequencing of a normalized and diverse cDNA pool, the issue of unequal transcript abundance remains and will likely require at least $15\times$ coverage or potentially more to obtain a broad, full length assembly of a transcriptome pool.

Transcriptome assembly

454 sequence data has been used extensively for de novo assembly of bacterial genomes, where it performs very well although repeat regions larger than the sequence read lengths are problematic (Holt and Jones 2008; Pop and Salzberg 2008). Here we assess the ability to assemble a transcriptome from 454 EST sequences de novo (i.e. without the aid of Sanger EST sequences or WGS). This is important as ecological systems generally do not have the molecular resources available for aiding contig assembly. Determining which EST sequences, of the hundreds of thousands of short sequence reads, belong together and should be assembled as a contig representing the cDNA they were shotgun sequenced from is computationally demanding but relatively easy with a reference genome. For example, 454 sequencing was used to sequence random portions of the Neanderthal genome, which was then readily aligned and compared with the human WGS (Noonan et al. 2006).

Contig assembly programs designed for the quality scores and longer sequence reads of Sanger sequence data perform poorly with the shorter 454 ESTs (Chaisson et al. 2004; Pop and Salzberg 2008). However, even the specially designed, commercial software supplied with 454 sequencers (Newbler assembler), which uses the “flowgram signal space” information unique to 454 sequencing (454 Life Sciences, Inc), also seems to perform poorly with de novo transcriptome assembly although perhaps better than other programs (Weber et al. 2007).

De novo transcriptome assembly from a single run of normalized mRNA pooled from 4 tissues of the Barrel clover plant (*Medicago truncatula*) was poor (e.g. only two contig assemblies over 400 bp in length; Cheung et al. 2006). Estimated transcriptome coverage of the 454

sequencing run in the *M. truncatula* study was very low ($0.28\times$). Thus the probability of having multiple overlapping sequences was low and likely detrimentally affected assembly. A separate assembly was attempted on mRNA from another plant, *Arabidopsis thaliana*. In this study, unnormalized cDNA was sequenced using two 454 runs on only a single tissue type (8 day old seedlings; Weber et al. 2007). Although average coverage is difficult to estimate given that unnormalized mRNA was sequenced, compared to the previous study transcript diversity has been reduced while coverage has increased. This resulted in at least 10,000 of the 17,449 genes found being covered by at least 3 ESTs. The performance of three different contig assembly methods (Newbler, CAP3, and stackPACK EST analysis pipeline) were compared, with very few full length cDNA contigs assembled even when there was sufficient EST coverage across the cDNA in question (Weber et al. 2007). Thus, even when transcriptome coverage is good, de novo transcriptome assembly appeared to be a difficult goal to attain due to the poor performance of Sanger assembly algorithms. However, Vera et al. (2007) have recently had good success with de novo transcriptome assembly of 454 data and their findings discussed below shed light on previous difficulties.

In order to develop genomic tools in the Glanville fritillary butterfly, a model ecological system with essentially no genomic sequence, 454 sequencing was used to rapidly characterize its transcriptome (Vera et al. 2007). Using a normalized cDNA pool, derived from diverse mRNA pools, two 454 runs were performed. After filtering out low quality sequence and amplification primers, assembly was performed using the commercial software program SeqmanPro v.7.1, of the Lasergene software package, which can be specially parameterized for 454 sequence and incorporate quality scores of the “flowgram signal space” during short read assembly (Lasergene Inc.). After filtering, a total of 608,053 ESTs (mean length = 110 bp) remained which assembled into 48,354 contigs and 59,943 remaining singletons. Singletons are individual ESTs that were not able to be grouped into a contig, but were still high quality sequence reads. The longest 4800 contigs ranged in length between 348 to 2849 bp and had an average coverage level of 6.5 EST reads.

Assembled contigs were compared with 3,888 Sanger sequenced ESTs to assess 454 error rates in this dataset. 749 Sanger contigs and singletons found matches with the 454 contigs and were 97% identical, with a 454 gap rate of ~ 3 per 1000 bp (Vera et al. 2007). This is likely an overestimate of the sequencing error rate due to polymorphism in both datasets and analysis of regions having low sequence coverage. Average transcriptome coverage depth was estimated at $2.3\times$, based on assuming a similar number of genes as the GRS *Bombyx mori* (18,000) and an

average gene length of 1500 bp (Vera et al. 2007). Thus, increased EST sequence coverage and a new assembly method resulted in a significantly improved de novo assembly compared to the two previous attempts in plants presented above. Note however that this level of coverage is much less than the 15× coverage recommended in earlier sections, and a lot less than the bacterial genome assembly work which had ~40× coverage (Margulies et al. 2005). Due to this low coverage, the full transcriptome was not assembled, which may be a difficult goal to obtain given transcriptome complexity (e.g. alternative splicing discussed two paragraphs below).

How does one assess the quality of a de novo transcriptome assembly of a species without genomic resources? Again, we turn to the butterfly 454 contig assembly as an exemplar, where performance was assessed in several ways. First, estimation of the 454 error rate discussed in the previous section began with comparing long, Sanger generated cDNA sequence with the assembled 454 contigs, revealing accurate and long 454 contigs (Vera et al. 2007). Performance of the assembly was further assessed by comparing the relationship between 454 EST sequence coverage depth and length. As the number of EST sequences for a given cDNA increases, contig assembly length should also increase. Solely comparing this data with itself says nothing about the accuracy of the assembly, as assembled contigs could be junk (e.g. repetitive DNA). However, by comparing the depth and length of contig coverage with reference to orthologous genes in a GRS, one is able to visualize whether increased sequence depth results in increase length coverage of cDNA. Such a relationship is clearly visible when this comparison was performed by Vera et al (2007), showing that short cDNA genes are fully covered and longer cDNAs have a linear relationship between coverage depth and length (Fig. 2).

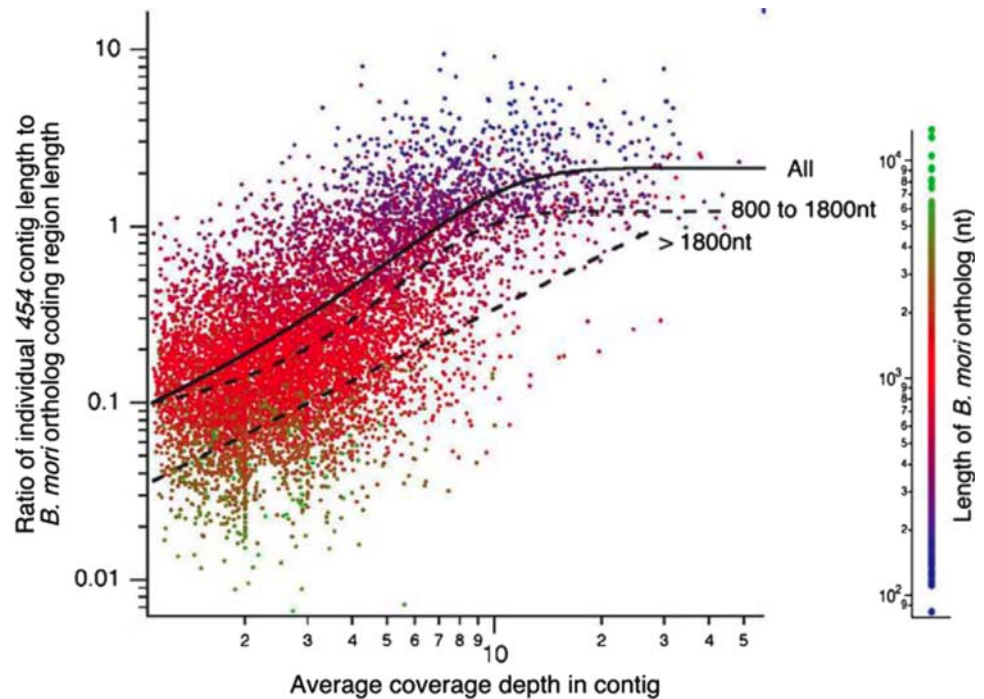
Assembly of a large number of sequence reads from a limited and normalized mRNA pool results in deep sequence coverage with several important outcomes. The first consequence of deep EST coverage emerges from a comparison of the coverage between traditional Sanger sequencing and 454 sequencing of the same normalized cDNA material. Sanger sequencing coverage is ideally very thin (i.e. <2 reads per unigene; a unigene is a unique gene) to maximize gene discovery while 454 contigs assembled from sequencing such cDNA pools have many regions of overlap. As mentioned above, this greater overlap can provide a greater level of accuracy (Goldberg et al. 2006; Margulies et al. 2005; Moore et al. 2006; Wicker et al. 2006). Second, deep coverage also finds many alternative splicing variants of many different transcripts. While a gene of 10 exons would normally have mRNA consisting of exons 1–10 connected together,

alternative splicing can result in permutations of exclusion, producing different mRNA transcripts (e.g. mRNA with exons 3–5 excluded, producing only exons 1–2, 6–10 connected together). The importance of alternative splicing variation to the study of ecological genetics is only recently beginning to be appreciated (Marden 2006). While interesting, variable alternative splicing products can cause severe assembly problems and thus need to be identified and treated separately during assembly (Vera et al. 2007). Again, alternative splicing may be a significant impediment to full de novo transcriptome assembly and warrants further study.

Deep coverage also provides SNP data. The two plant example reviewed earlier that attempted de novo transcriptome assembly used mRNA from a very homogeneous genome (e.g. Weber et al. isolated their mRNA from a single accession line of the selfing plant *A. thaliana*; Weber et al. 2007). In contrast, Vera et al. (2007) used various tissue and developmental stages from ~80 individuals from eight families of an outcrossing diploid butterfly. Thus, Vera et al. (2007) necessarily had a more complex mRNA pool due to diverse transcripts and genetic polymorphism. However, this diverse pool was collected on purpose since contig assembly brings together sequence reads of the same genetic region from many different individuals. Initial concerns regarding the effect of natural genetic variation on assembly appear to be unwarranted, at least with the level of population diversity sampled (Vera et al. 2007). With aligned sequences of the same region, observed SNPs can be verified as not being sequencing errors by identifying multiple independent reads of the alternative bps comprising a given SNP. We will return to the various uses of mRNA pooling from individuals for finding polymorphism in the SNP section.

Assembly results such as those obtained by Vera et al. (2007) should be attainable with any organism's transcriptome given proper depth of coverage and assembly method. Moreover, given the increasing read lengths of the new GS FLX runs, advances in chemistry and error calling, assembly using such data should be of a generally higher quality in terms of longer contig assembly length, lower error rate, and generally a greater transcriptome coverage. Here it should be noted that little if any of the bioinformatics community's attention is focused on developing software dedicated specifically to the problem of de novo transcriptome assembly from short read sequence. Indeed recent reviews on the bioinformatics challenges presented by short read sequence, and the resulting software developments, do not even consider de novo transcriptome assembly challenges (Holt and Jones 2008; Pop and Salzberg 2008).

Fig. 2 Relationship between Glanville fritillary 454 EST sequence coverage depth and length, with reference to cDNA of the genomic reference species *B. mori*. X axis shows increasing number of Glanville fritillary 454 ESTs per contig, while Y axis shows percent coverage by Glanville fritillary contigs of full length *B. mori* orthologous genes. Figure from Vera et al. (2007)



Annotation & assessing transcriptome coverage

What can ecologists hope to do with all this data? This depends on the types of questions that are being asked. Keep in mind that one 454 run will likely generate ~300,000 sequence reads which may assemble into ~50,000 contigs with equal or more singletons. Such datasets can be downloaded from the recently launched short read archive of trace data for next generation sequencing projects (<http://www.ncbi.nlm.nih.gov/Traces/sra>). At one extreme, when the goal is to design a microarray to assess global expression in relation to a given phenotype polymorphism, then knowing how many genes were found and what they are is vital. Knowing you have only covered 20% vs. 70% of your transcriptome is also important, since you might want to invest more money in better coverage to increase the chances of finding important genes in planned microarray studies. Researchers solely using 454 sequencing to generate molecular markers will find the section dedicated to those more informative. Here the issues involved in annotation and transcriptome coverage assessment are discussed.

The first step is to determine what type of data one has through comparisons with annotated sequence databases. Through such comparisons, one can begin to attach names to genes and infer function. This annotation of the assembled transcriptome can be difficult, not conceptually but rather computationally, given the extremely large datafiles and searching requirements, results storage, and later retrieval of results in a easy, quick, meaningful fashion. In sum, a non-model ecology lab will need to develop

database management skills (e.g. Papanicolaou et al. 2008; Paschall et al. 2004).

Open source database programs are available, such as MySQL, which are able to manage the large number of reference spreadsheets, storing information from various database searches, etc. Mining of this database can be accomplished using simple search commands. Sequence annotation to data mining pipelines are in a period of flux, with many different labs reinventing the wheel for their particular research programs. A pipeline refers to a series of programs that take an input set of data, hand it off from one program to another, and then output a finished product. As such, pipeline in this context would take raw 454 flowgram data, identify high quality sequence and trim away poor quality data, assemble this into contigs, annotate these contigs through comparison with sequence databases, identify SNPs, etc., and output this in a simple format for someone to use.

While some pipelines, or packages, are starting to emerge as viable options for use in this context (e.g. Open Sputnik Beldade et al. 2006; Funnybase Paschall et al. 2004), this field is awash in many labs building separate pipelines for their own needs. For genomic data, the Generic Model Organism Database (GMOD) project appears to be a future leader due to National Institutes of Health and most recently, National Evolutionary Synthesis Center (NESCent) support (e.g. the first GMOD summer school was held in summer 2008). GMOD is a collection of interoperable open source software components for managing genomic data (Stein et al. 2002). Such a project needs to emerge for handling the EST data generated from

the current and growing number of labs for both Sanger and next generation sequencing, providing a centralized grouping of open source software that could be continually improved over time. Importantly, this would create tools that are readily accessible for old and new labs alike, which would lower the barriers to getting funding when working with next generation sequence data. Being able to integrate this data with GMOD components should be a longterm goal, using the GRS concept as a bridge for generating candidate linkage assignments which could be annotated by various labs across different taxa. The amount of investment needed in this aspect of data management should not be underestimated. Simply put, these will be the tools that allow you to actually use the data you generate. Thankfully, intensive courses and conferences are emerging to directly deal with these issues and collaboration with bioinformatics specialists is the norm.

We will start assessing the transcriptome coverage of 454 sequencing by further reviewing the excellent *A. thaliana* paper discussed earlier. Weber et al. (2007) explicitly used two 454 sequencing runs to characterize unnormalized mRNA from 8 days old seedlings, as transcripts expressed in these tissues had been previously well characterized using microarrays and the 454 data could be directly compared with this as well as WGS data (Schmid et al. 2005b). The unnormalized mRNA pool thus contained a high variance in mRNA copy number among genes, with nearly 3000 genes having only one 454 EST each and a handful of others having over four orders of magnitude more 454 ESTs. This small number of highly expressed genes represent 26% of all the 454 ESTs matching the *A. thaliana* transcriptome ($n = \sim 541,000$). For example, 5 genes for Rubisco had more than 85,000 ESTs, while those for the 20 genes encoding chlorophyll *a/b*-binding proteins had approx. 60,000 ESTs (Weber et al. 2007).

Weber et al.'s (2007) first 454 sequencing run found 59% of predicted *A. thaliana* gene models (transcribed genes and alternatively spliced variants), a level expected as only 55–67% of *A. thaliana* genes are expressed in a single organ (Schmid et al. 2005b). Consistent with finding nearly all of the unique transcripts in the mRNA pool, a second 454 run only produced 10% more genes, while increasing cDNA sequence coverage by 50% (from 7 to 10.3 Mb; (Weber et al. 2007). The authors concluded that their two runs detected “at least 90% of all genes expressed in this sample”, including those with low expression levels (Weber et al. 2007).

Weber et al (2007) next assessed whether there was any bias in the shearing of the cDNA during sample preparation which could have resulted in uneven 454 EST coverage of cDNAs, by looking to see where the 454 EST actually were relative to the known cDNA. This was relatively easy given the extensive genomic resources for *A. thaliana*, which has

$\sim 700,000$ Sanger ESTs and well annotated WGS. They found that all regions of the cDNAs were covered by the 454 ESTs with low to moderate 5', middle, or 3' coverage biases for different cDNA size and expression classes, concluding that EST coverage was sufficient for contig assembly (Weber et al. 2007).

Interestingly, 454 sequencing identified sequences that matched genome sequence but were not found in the large Sanger EST collection, with an additional 648 gene loci identified and 5302 known gene loci provided with additional length coverage. The authors state, “it is likely that some of the 648 loci detected by pyrosequencing, but not in (the $\sim 700,000$ Sanger ESTs), represent difficult to clone sequence or DNA molecules that are toxic or otherwise unstable in *E. coli*” (Weber et al. 2007). In sum, of the 23,367 genes in *A. thaliana* (many of which have several gene models), 74% were at least partially covered by two runs of 454 sequencing of an unnormalized library from one tissue type.

These authors also wanted to determine the relative performance of 454 vs. Sanger sequencing in finding novel genes. Since Sanger sequencing 10,000 ESTs costs roughly the equivalent of two 454 sequencing runs, the authors went on to generate 5 random sets of 10,000 ESTs from the *A. thaliana* EST database (containing ~ 700 k Sanger sequences) and found that these random sets on average only find a third ($n = \sim 5500$) of the genes recovered by the 454 sequencing effort ($n = 17,449$).

Let us now consider how to assess transcriptome coverage for a species which has little to no genomic resources. We must turn to blast searches against sequence databases. Luckily, many genes are shared across divergent taxa and will be readily identified as such in blast databases searches (e.g. housekeeping genes evolving under strong functional constraint such as central metabolic enzymes). Many genes are also likely to find an uninformative homologous blast hit in reference databases (e.g. gene descriptions of “hypothetical protein” are common), but at least one knows that such a contig is a likely coding gene. Only 30–40% of unigenes from relatively large cDNA libraries usually annotate as such, with the majority having no significant hit above a given quality threshold (i.e. low quality hit scores could happen by chance; e.g. Beldade et al. 2006; Paschall et al. 2004).

Vera et al. (2007) used this approach to estimate their transcriptome coverage, blasting their butterfly 454 contigs against several relevant databases and then summing all the different unigenes that could be identified. This allowed for multiple contigs to hit a given unigene (Paschall et al. 2004), as the fragmented nature of 454 sequencing produced many non-overlapping contigs for the same full length sequence (Fig. 1). Evidence for 9311 unigenes were found, representing $\sim 50\%$ of the genes estimated to be in

the closest GRS, the silk moth, *B. mori* (total WGS predicted genes = 18,510; Xia et al. 2004). Certainly this approach misses genes that have become highly diverged and diversified since butterflies and moths last shared an ancestor, but it does provide a robust lower bound on coverage.

Here I present a way to estimate a potential upper bound of transcriptome coverage via comparative coverage inference (H. Vogel personal communication). House-keeping genes have three properties ideal for this task. First, they evolve under strong purifying selection and are thus easily annotated across divergent taxa. Second, specific subsets of them are highly expressed in nearly all tissues, such as proteins in the ribosome and oxidative phosphorylation. Third, bioinformatics tools have evolved to a current state of being easy to use and highly informative, with the KAAS web interface being particularly good (Moriya et al. 2007). Comparative coverage inference rests on the assumptions of random sequencing from a normalized cDNA pool and a relatively similar numbers of genes between the two species being compared. With these assumptions satisfied, or mostly satisfied, one can compare the numbers of genes hit within genetic pathways between a GRS and a focal species. GRS provide a full, or nearly full, set of genes predicted from WGS and therefore a more complete set of genes compared to gene sets derived from sequenced cDNA pools such as Sanger ESTs.

Using the butterfly 454 data of Vera et al. (2007), normalization of the cDNA pool was assessed by comparing the average number of ESTs per ribosomal protein contig to the average number of ESTs per contig for the dataset as a whole (for contigs with blast hit bit scores ≥ 45). Ribosomal proteins have an average of 37 ESTs per contig. Importantly, the 90th quantile of ribosomal protein ESTs per contig ($n = 111$) is nearly 5 times that of the average number of ESTs per contig for the rest of the data ($n = 22$). Normalization is expected to reduce the high variance among genes in mRNA copy number, exemplified in the *A. thaliana* study above, to within a 10 \times difference among genes (Bonaldo et al. 1996). A successful normalization would therefore result in the majority of mRNAs (e.g. 90th quantile) from a highly expressed gene set being within 5 \times of the mean for the rest of the sample. Thus, the normalization was successful.

We can now ask what are the relative numbers of genes hit in different, less expressed pathways for the assembled butterfly transcriptome compared to WGS predicted genes in the GRS (*B. mori*). Across five different central metabolic pathways, the assembled butterfly 454 contigs hit, on average, 70% of GRS genes (Fig. 3). This provides an upper estimate of transcriptome coverage that is a 20% increase over the lower estimate (i.e. a potential upper limit of 13,142 genes based on GRS inference).

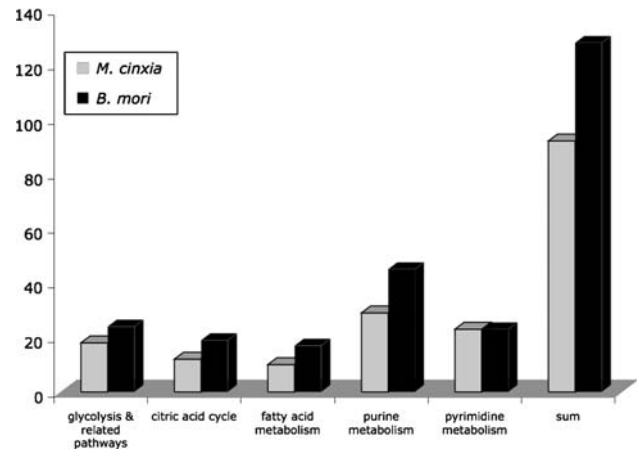


Fig. 3 Number of genes (Y axis) per metabolic pathway (X axis) identified from the assembled *M. cinxia* transcriptome compared with the WGS predicted genes of *B. mori*. Analysis used the web interface KAAS (Moriya et al. 2007)

As a final note, this assessment of transcriptome coverage assumes that all unique transcripts were in the original mRNA pool that was sequenced, which is certainly not the case. Many developmental, immune, or environmentally induced genes require specific conditions for expression and thus may be absent from such mRNA pools, even when many developmental stages and tissues are purposefully included. However, such genes are not likely to be a substantial fraction of the transcriptome. Yet, with this issue in mind the importance of the research questions to ultimately be addressed cannot be overstated, as questions should guide the type of tissues used and treatments applied to the live organism from which the mRNA will be isolated. For example, research programs interested in ecological immunity should certainly infect some of the individuals used to make the mRNA pools for 454 sequencing. Should a research program wish to gain access to the as much of the transcriptome as possible to provide as a community resource, collecting mRNA from varying tissues and developmental stages from individuals which underwent a series of inductions and treatments would be ideal.

Microarrays

Microarrays are generally used to determine the relative gene expression between two or more groups on individuals for as many genes as can fit on a microscope slide. For the molecular ecologist, this tool allows one to query the transcriptome for genes that might be differentially expressed between two or more groups in an effort to identify candidate genes for further study (e.g. Abzhanov

et al. 2006). Currently there are many different technologies for microarray construction and they warrant careful consideration when planning such work. Here I will focus on what might be referred to as classical microarrays, which are simply glass slides with very small dots of DNA attached to them, to which other DNA or RNA can hybridize in a dot specific manner. Hybridizing material is labeled with fluorescent dyes and hybridized to the DNA dots on the slide, which after washing can be read with a very precise laser scanner at several different wavelengths, allowing the relative intensities of each dye by dot to be measured. This allows for a rapid, gene (dot) specific quantification of relative amounts of hybridizing material, which is generally some form of RNA or cDNA. There are many different ways of constructing, using, and analyzing microarrays, with final design generally being an informed compromise between biological question, analysis power, and costs (e.g. Shiu and Borevitz 2008).

As an exemplar route to designing a microarray from 454 transcriptome data, the work of Vera et al. (2007) is further reviewed. An open question was whether the short contigs generated by 454 sequencing could provide enough sequence to design microarray probes that performed well. To assess this, assembled 454 contigs and singletons were selected and six probes designed for each using Agilent's eArray web tool (<http://earray.chem.agilent.com/earray/>).

Each probe (60 bp in length), corresponding to a stretch of DNA which was selected by calculated hybridization properties of this stretch of DNA, was printed on a glass slide containing 244,000 features (dots of probes) and hybridized with dye labeled RNA from the pool used for the 454 sequencing. Of the six probes per contig, one was selected based on optimal performance and greatest 3' proximity in the contig (contig orientation was determined from blast inferred open reading frame against a GRS, when possible). Probe performance could be an assessment of relative intensity compared with negative controls and/or lowest variance across repeated hybridizations (Vera et al. 2007 used only the former). Probe proximity to the 3' end of the transcript was used as an additional criteria as mRNA degradation proceeds from the 5' to 3' end (Lee et al. 2005). This validation of probe performance is an important step, ensuring that the final selected probe for each contig performs well, at least with the test sample material.

At this point it is worth reiterating that the process described above (i.e. constructing a custom, richly featured microarray) did not involve any handling of library clones or benchwork of any kind. Rather, it was all done from a computer using assembled and annotated 454 sequence data, ported to a free web-based tool for oligonucleotide probe design, random array layout, and printing. This is substantially faster, cheaper, and less error-prone than

previous methods of constructing cDNA arrays by spotting PCR products from library clones. Of course, the sequence data generated from any source can be used to make such microarrays. For example, if you wanted to make a microarray today of your favorite set of genes for your GRS, or any species with DNA data such as a cDNA library database, you could simply use this data on the eArray web tool. In fact, using data from most cDNA databases would even tell you the orientation of the cDNA (since cDNA libraries are directionally cloned and sequenced). All one needs at this stage is the DNA data. The cost of purchasing such commercially printed arrays is much higher, which somewhat negates the funds saved during construction, but the savings in time and error reduction remain and should not be underestimated.

Final array design used 13,780 validated probes across roughly 9–13,000 different genes (lower and upper unigene estimates). Each probe was printed in at least triplicate (genes of interest were printed 5 times) randomly across the array in a 44,000 feature array format, with four arrays per slide. Replicate hybridization across arrays revealed excellent repeatability indicating low technical error, which is likely due to using validated probes and high quality arrays (Vera et al. 2007). Of course, Agilent is not the only company providing high quality microarrays, but simply the one used by Vera et al. (2007), because of the ease of custom array design and high quality of final printed slides. Optimal experimental design and protocol will depend on particular hypotheses and access to local experience in using microarrays. Slide performance and costs are changing rapidly, with commercial custom designed slides becoming much cheaper with increased performance. The technological landscape is changing faster today than at any time in the past, and becoming well versed in the latest techniques and companies offering them is an important step in genomic tool development and proper cost benefit analysis of funded objectives. These examples here should be considered as providing a proof of concept rather than an exact roadmap.

A final consideration is that since 454 sequencing runs directly sequence individual mRNAs within a sample, it can provide a measure of the absolute number of different transcripts. This quantification of mRNA level across genes is similar to analyses which measure relative hybridization intensity among genes via microarrays. Comparisons between the number of 454 ESTs per locus with microarray results for *A. thaliana* showed a correlation coefficient of 0.45 (Weber et al. 2007), while comparisons across 454 runs in *Drosophila melanogaster* were much higher, with correlation coefficients of 0.83–0.91, which is similar to replicate microarray experiments (Torres et al. 2008). Interestingly, in the Weber et al. (2007) study of *A. thaliana*, their 454 data also showed expression levels

for a number of genes which were not on the commercially designed microarrays ($n = 1410$). Both methods have their specific biases, with microarrays having biases due to dyes, hybridization performance, etc., while biases in 454 sequencing, other than the errors discussed earlier, arise due to unequal coverage of variable length cDNA, such that very short or very long cDNAs are under represented, which can result in a significant bias regarding expression level determinations (Torres et al. 2008). Shearing of cDNA such that all genes have similar sizes is likely to alleviate this bias and result in similar expression insights between 454 and different microarray platforms (Torres et al. 2008). Finally, unlike traditional microarrays, using 454 for expression profiling can be done on any species and mapped to genomic regions for species having WGS, or a suitable GRS, highlighting the power and promise of massively parallel direct sequencing to extend biological insight beyond traditional methods.

Molecular markers: SNPs, microsatellites, and EPICs

Polymorphic markers within and among individuals have long been used in molecular ecology for assessing relatedness, demographic structure and history, and identification of candidate loci underlying phenotypes of interest (Luikart et al. 2003). This latter issue is further discussed in the following section on *Genomic scans*. The relative performance of SNPs versus microsatellites in these contexts has been well reviewed elsewhere (Morin et al. 2004; Zhang and Hewitt 2003), concluding that in general 2–6 times as many SNPs are needed compared to polymorphic microsatellite loci. This is due to the general limit of two alternative states for SNPs compared to the much greater number of states at a given microsatellite locus.

The specifics of which marker to use and their relative performance depends greatly upon the questions being addressed and the system being studied. However, SNPs in general provide higher genotyping efficiency, data quality, genomic coverage, and low probability of homoplasy (Morin et al. 2004). However, one serious concern with the use of SNPs is ascertainment bias (Clark et al. 2005). Ascertainment bias arises due to selecting SNPs from a non-representative, usually small sample, where high frequency SNP alleles are found more readily than rare ones. Subsequently taking these SNP loci and measuring their frequencies in larger field samples results in only studying common alleles and thus population genetic measures depending on frequency information are biased, such as estimates of nucleotide diversity, Tajima'D, Fst, and linkage disequilibrium (Clark et al. 2005). Appropriate

research design and incorporation of correction methods are thus necessary (Rosenblum and Novembre 2007).

Planned sampling of individuals for 454 sequencing provides an excellent opportunity to generate large numbers of both SNP and microsatellite markers. While previous studies have used 454 sequencing for finding large numbers of SNPs, this was performed using WGS as a reference for SNP determination (Barbazuk et al. 2007). However, developing such markers is common in systems without genomic resources (Morin et al. 2004). For example, Beldade et al. (2006) used over 20 outbred butterfly (*Bicyclus anynana*) individuals in their wing tissue cDNA library and were able to identify 320 candidate microsatellite loci and over 14,000 candidate SNPs, with at least 316 genes identified as having at least one high confidence SNPs among their 9,900 Sanger EST sequences.

With a greater depth and breadth of coverage, 454 sequencing outperforms traditional methods in molecular marker identification. Using the 454 sequence data from the Vera et al. (2007) butterfly study, which was derived from ~80 individuals from eight families, a quick scan of the assembled contigs finds a total of 1063 candidate microsatellite loci across di-, tri-, tetra-, and pentanucleotide repeats (Table 1). These candidate marker loci are likely to be fundamentally different to previous microsatellites generated for Lepidoptera, as they are associated directly with coding genes instead of, like many Lepidopteran microsatellites, being located in repetitive DNA regions (Van't Hof et al. 2007). Recent efforts in the same butterfly species (*Melitaea cinxia*) to find new microsatellites via screening 4 DNA libraries only yielded 37 candidate loci of which 5 showed polymorphism (Sarhan 2006). Scanning of the 454 data for SNPs identifies more than 2000 contigs having at least one high quality SNP (C. Vera unpublished data), with an average SNP density of 6.7 SNPs per 1000 bp of coding DNA, which is similar to other species (Morin et al. 2004; Wondji et al. 2007). When these SNPs are located in contigs that have annotation data, the SNP data can be further dissected to obtain information regarding location (1st/2nd/3rd position of a codon or UTR region) (Fig. 4).

Table 1 Candidate microsatellite from repeat scan of 454 contigs using Sputnik repeat finder (<http://cbl.labri.u-bordeaux.fr/outils/Pise/sputnik.html>)

Microsatellite repeat	Loci
2	139
3	376
4	276
5	272
Total	1063

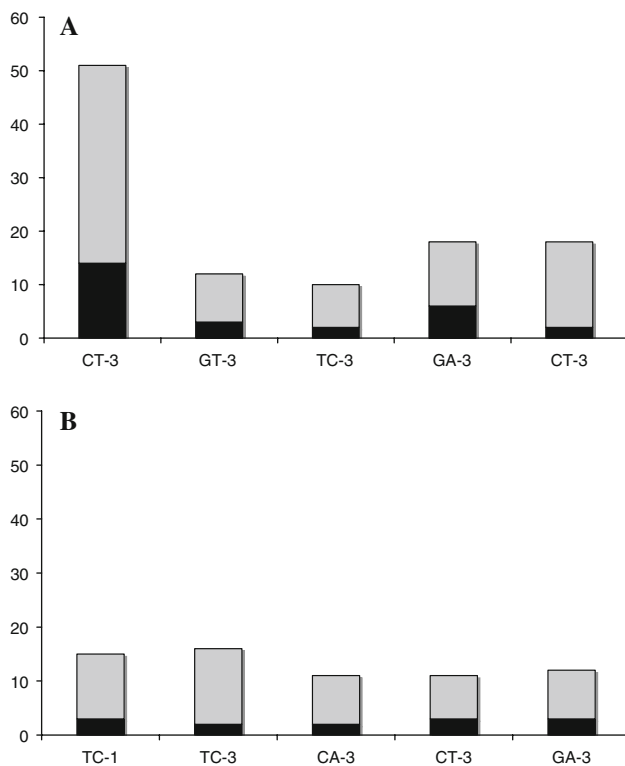


Fig. 4 Representative SNPs for Glanville fritillary contigs with hits to the genes (a) Spodopocin (best blast bitscore = 120) and (b) ribosomal protein L40 (best blast bitscore = 303). Histograms show total EST depth of coverage at each SNP, and relative frequency of the two alternative nucleotides at each position are indicated by shading. X axis shows SNP nucleotides and codon position. For example, SNP TC-1 is the 1st position of a codon in ribosomal protein L40 which codes for a leucine, which has a degenerate first position

Additionally, one can use the deep coverage per contig to capture specific aspects of SNPs, such as their relative frequency within a deeply covered region (Fig. 4). However, keep in mind that such insights require adequate EST coverage which likely entails a simple mRNA source or at least 2 if not more 454 sequencing runs for a more complex source (at least with the pre-GS FLX technology). One can then query a database for SNPs that are found, for example, at least two times at a given nucleotide position, when the depth of coverage is equal to or greater than 10 ESTs. This provides for a means of ensuring SNPs are accurate and if the frequency is recorded, an initial estimate of SNP frequency within the pooled mRNA sample. Ultimately, one could query for high frequency SNPs likely to be most informative in paternity assignment studies (Morin et al. 2004), such as those in silent sites which have at least a 40% frequency with an EST depth of 10 or greater. Others have used such SNP frequency information to estimate the mutation rates of Sanger sequencing and cDNA library construction, as well as population levels of heterozygosity (Long et al. 2007).

SNP and microsatellite marker classes can also be integrated together in the form of what is called SNP-STRs, having a SNP within several hundred bp of a microsatellite. In humans, SNPs have an average mutation rate of $\sim 2.5 \times 10^{-8}$, which is much slower than microsatellites (10^{-2} – 10^{-5}) (Tishkoff and Verrelli 2003), and thus together they provide a dual level of temporal inference excellent for inferring recent evolutionary events (Agrafioti and Stumpf 2007; Mountain et al. 2002). For example, SNP-STRs have been developed and used to study the timing of African cichlid fish radiations (Won et al. 2005). However, in *D. melanogaster* these two mutational classes have similar rates and therefore provide similar temporal insights (Storz 2005).

How does research proceed when the gene or genes of interest have no SNP or microsatellite variation? Such a situation might face a lab that only performed a partial or single 454 sequencing run on their material, had a bottlenecked population, incorrectly calculated their projected sequencing coverage, or were just unlucky. Under these situations 454 sequencing is not likely to find an abundance of SNPs and perhaps not that many microsatellites. However, this is hope, EPIC hope. Exon priming intron crossing (EPIC) markers are PCR-based markers that have primers located in exons, but have amplicons that include introns when genomic DNA is used as template. By using a GRS, one can quickly take cDNA contigs and predict intron locations for making EPIC markers (Bouck and Vision 2007). Introns have much higher rates of diversification compared to coding genes, containing many SNPs and indels, and are thus make a suitable marker for many of the aforementioned studies (Bouck and Vision 2007; Zhang and Hewitt 2003).

Candidate genes

Many ecologists wishing to develop a functional genomics approach in their non-model system have delved into the literature enough to determine that in other, perhaps genomic model species, the genes or genetic pathways affecting their phenotype of interest are well understood. In such cases, assuming a similar genetic architecture can be a great starting point for study (e.g. Hanski and Saccheri 2006; Nachman et al. 2003). Thus, many researchers have candidate genes for which they wish to quantify coding and expression variation (Ellegren and Sheldon 2008). While the previous section on microarrays addressed the development of a tool for large scale (i.e. many gene) expression measurements, here we consider the development of genetic sequence for a targeted expression and coding sequence study. Again, the butterfly system presented

above provides a good example of a candidate gene approach in the 454 context.

Previous work in the Glanville fritillary suggested that dispersal within its metapopulation in Finland has a genetic basis (Hanski et al. 2004; Saastamoinen 2007). Exposure to evolutionary genetic study in *Colias* butterflies, a long running research system which has build a strong case for single gene performance and fitness consequences, provided a candidate gene for study called phosphoglucose isomerase (Pgi) (Watt 2003; Wheat et al. 2006). Pgi catalyzes the second step of glycolysis and is highly polymorphic in the Glanville fritillary (Haag et al. 2005). Initial study of *Pgi* variation in the Glanville fritillary found significant allelic association with flight performance and fecundity fitness measures (Haag et al. 2005; Saastamoinen and Hanski 2008), consistent with the previous work in the nearly 80 million year distant *Colias* butterflies (Braby and Trueman 2006; Wheat et al. 2007). Thus, the enzymes of glycolysis in general, and *Pgi* in particular, were of special interest when examining 454 sequencing results.

Partial sequence of 9 of the 10 genes of glycolysis are found in the assembled 454 contigs (hexokinase was the exception; it was absent in the 454 data, has little or no representation in other butterfly cDNA collections and thus appears to be a very low expression gene in Lepidoptera). There are several independent, non-overlapping contigs for many of these genes. For traditional cloned cDNA libraries, the individual clone can be taken out of the freezer and the entire cDNA easily sequenced. When working with 454 data, there is no clone to return to, which highlights one of the most fundamental differences between these two methods.

For each of these 9 genes full length sequence is desired. There are standard methods for getting full length cDNA sequence starting from an initial fragment of the cDNA, such as using rapid amplification of cDNA ends (RACE) techniques (Zhang and Austin 1997). However, these do not always work and cloning your favorite gene or set of genes can run into difficulties which might require more than a trivial investment of time (e.g. SMARTTM RACE cDNA Amplification Kit, Clonteh, Inc.). Once full length sequence is obtained, primers can be designed in the 5' and 3' flanking regions for whole gene sequencing from population samples to determine the extent and distribution of nucleotide variation in either cDNA or entire genomic regions including introns (e.g. Wheat et al. 2006).

Partial sequences can also be used for mRNA quantification. Recall that many of the probes put onto the microarray of Vera et al. were singletons, consisting of <200 bps, which nevertheless hybridized well (Vera et al. 2007). Such fragments can also be used for designing real time PCR studies for independent

quantification of gene specific mRNA levels, as the optimal PCR amplicon lengths for real time PCR are ~50–150 bp. Using a GRS to infer gene exon/intron boundaries, primers for real time PCR can be designed that span an exon/intron boundary which is an ideal way to minimize potential genomic DNA contamination during amplification. Importantly, the 454 contigs also provide a vast resource for selecting control genes to use in such expression studies.

Genomic scans

With 100's to 1000's of molecular markers spread out across the genome, it is possible to start using them for genomic scans for selection (for more detailed information see the excellent reviews by Luikart et al. 2003; Storz 2005). These are powerful methods that can well complement ecological genetic study and 454 sequencing has the power to put more of these markers, more quickly into the hands of research teams who have the field systems to exploit their power, as compared to traditional Sanger sequencing of cDNA libraries (Bouck and Vision 2007). These genomic scan methods, simply put, compare between or among populations for large sets of markers and are expected to find similar levels of diversity within and between populations across all markers, while markers in or near genetic regions affected by selection should be outliers having too much or little divergence or diversity (Storz 2005). Knowing the relative placement of markers along chromosomes can help in such analysis, as flanking markers are likely to show similar patterns and thus reinforce trends (e.g. Nair et al. 2003). Model-based and model-free tests can be used, with the former dependent upon null distributions given simulated datasets of differing demography and the latter using empirically derived expectations, both of which have their strengths and weaknesses (Storz 2005).

Genome wide estimates of diversity are starting to emerge from Sanger transcriptome sequencing based on SNP variation (Long et al. 2007). However, the deep coverage afforded by 454 sequencing allows for greater SNP estimates for a greater number of loci than afforded by previous Sanger estimates. With adequate sampling within a population, this deep coverage per locus should provide relatively robust genome diversity estimates, although this depends upon the underlying population sampling. When coupled with the ability to include multiple individuals and populations into a single 454 run (Meyer et al. 2008), one could generate sufficient data for both *F_{st}*-based and diversity-ratio tests across loci, which differ in their demographic robustness and ability to identify candidate loci (Storz 2005).

What is the “proper” sampling per population? Again, research questions lead the way forward. Populations might represent two different pools of individuals of a given morphology type, or four different pools with two being an independent replicate of a given morphotype. Different species could even be used, allowing for a suite of additional molecular tests of identifying candidate genes under selection since the species last shared a common ancestor (e.g. MK and multilocus HKA tests; Nielsen 2001).

Genomic referencing summary

Throughout the various sections above the use of a GRS has been interwoven. Here I wish to revisit these various points to highlight their importance, for as more species have their genome sequenced, this concept will become increasingly relevant to future functional genomic studies on ecological model species. Although the day when WGS might be an option for many species is perhaps not too far off, having a group of people that are able to quickly bring together this data in any meaningful fashion will continue to be a limiting factor. Until this last barrier is removed, the ability to take sequence from any one species and infer meaningful genomic insights from another will be important.

Using Lepidoptera as a case example for this concept, there exists only one genomic reference species (*B. mori*), from which butterflies are approximately 100 million years divergent. Beyond providing an unbiased set of predicted genes, as genes predicted from WGS are independent of induced expression (unlike finding genes via transcriptome sequencing), GRS can potentially provide gene structure and order insights, although this needs to be tested in a given research system. Recently this concept was utilized to identify long exons (>500 bp) in *B. mori* WGS which could then be searched for in various butterfly cDNA libraries (Wahlberg and Wheat 2008). Having many such independent long exons is a long sought after resource for the development of phylogenomics. Comparisons of 10 genes across over 40 butterfly species, as well as other moths (i.e. Ditrysia), with the GRS *B. mori* indicates completely shared exon/intron boundaries across this evolutionary distance (~100 million years).

This important insight provides a means to rapidly, and on a genomic scale, design primers for SNPs identified in cDNA but which will work on genomic DNA samples (i.e. primers can be designed to PCR within exons, or across short introns, and not be located on exon/intron boundaries). Workflow for such an endeavor entails finding orthologous genes between the focal and GRS species, determining the exon/intron boundaries for the GRS species, and then using these coordinates to infer exon/intron boundaries in the focal species (Wahlberg and Wheat 2008).

Comparisons of genomic maps between butterflies and *B. mori* also indicate a relatively high level of shared gene order (synteny) (Yasukochi et al. 2006). This insight provides justification for having a candidate gene order derived from the GRS. When deciding which of 1000's of SNPs to focus upon, a candidate gene order or candidate genome distribution provides an opportunity to develop SNPs that have a potentially wide genomic distribution, concentration on a specific chromosome, or even within a chromosomal region. Synteny hypotheses can then be tested when research results suggests that such findings are of interest (i.e. investing the extra resources when necessary, rather than before). How would one use this? For example, with some level of synteny verified, SNPs can be selected for different levels of QTL study, from genome wide to local chromosomal regions. A word of caution is also needed here, for the rate of chromosomal rearrangements appear to vary dramatically across taxa and thus verification of synteny hypotheses are needed on a case by case basis (e.g. Bourque et al. 2005).

A further potential use of a GRS could be in contig assembly. Many of the contigs generated from both cDNA and 454 cDNA sequencing efforts contain only partial sequence of many different mRNA transcripts. As a result, several independent non-overlapping contigs can be formed for the same mRNA. One way of bringing these together is through comparison to a common set of reference genes. Using reference genes from a GRS makes the most sense for trying to group contigs for the same gene. In sum, a small investment in understanding the extent of shared genomic properties between a focal species and a GRS can have huge returns on future functional genomic study.

Drawbacks of 454 sequencing

There are several drawbacks of 454 sequencing that should be highlighted, lest they get lost in the details above. First, in the earlier section on error rates and elsewhere, the importance of generating the appropriate average coverage depths was emphasized. But again, the depth of coverage one needs to attain depends upon what one desires and can afford. Thin coverage over many genes will likely produce many false positive polymorphic sites, and as such, verification of these as SNPs would be a poor investment of time and money. However, bioinformatics advances have had a recent success using shallow coverage of genomic sequences, identifying and validating SNP with reference to WGS (Quinlan et al. 2008). While Vera et al. (2007) were able to assemble part of their transcriptome with deeper coverage, many genes nevertheless had limited and thin coverage. Again, while sufficient for microarray construction and generation of several 1000 high quality SNPs, not all genes of interest had high quality sequence and SNPs.

A second consideration is that the shotgun sequence data generated from highly fragmented cDNA is not always evenly distributed across cDNAs. Although Weber et al. (2007) reported a 5' bias in sequence coverage, they concluded that this was not very significant because on average there was even coverage across cDNA in their dataset as a whole. Vera et al. (2007) also found a similar, but not dramatic 5' bias, while Torres et al. (2008) report a substantial bias in which fragments were sequenced. Thus, terminal end and potentially fragment length biases appear to be common to 454 sequencing of transcriptomes, but the effect of these bias varies across studies depending on material and end use. With lower average sequence depth, this type of bias could potentially become more prominent, resulting in many genes potentially having only a limited amount of 5' coding sequence or fragment size information. Given the next generation of 454 sequencing, with improvements in sequence reads, length, and bioinformatics, this might not be a significant problem in the future. For example, paired end 454 sequencing provides a means of gaining data from the two opposing ends of each cDNA (Korbel et al. 2007), which simply circumvents any terminal end sequencing bias.

A third consideration is that once partial sequence from a gene of interest has been identified, getting the full length sequence can be difficult. Using traditional Sanger sequencing of a cDNA library would allow one to go back to a given cDNA clone and then fully sequence this clone. However, as noted earlier, such clones may not be full length unless special steps were taken during the construction of the cDNA library. Finding a gene fragment of interest in a 454 database would provide data suitable for using RACE techniques to obtain full length sequence. RACE would also have to be used in the event that a cDNA clone is not full length. Thus, obtaining full length sequence of a given gene via 454 sequencing compared to Sanger sequencing is a complicated issue which will vary on a gene by gene basis, depending upon sequencing coverage and quality of material used in both methods.

Conclusions

The speed and accuracy of next generation sequencing will continue to increase, driven primarily by biomedical and molecular biology interests which are already reaping huge rewards (Mardis 2008). This in turn will offer increased opportunities to develop functional genomic tools in many ecological model systems. The ability to use these next generation sequencing technologies for large scale sequencing of pooled individuals, and recover the information on an individual basis, is now available (Holt and Jones 2008; Meyer et al. 2008). This will be but one of the manifold impacts of 454 sequencing just now beginning to

be realized. However, the path to getting the right data in a timely fashion requires careful thinking and serious planning. Being able to both get the data and put it to work can be difficult, and thus collaboration across labs to build useful analysis pipelines are needed.

Thankfully, technological advances are making this process easier, which will allow scientists to focus on what they do best, posing interesting questions to push the boundaries of their fields forward towards more important and fundamental insights. However, the ability to pose questions that maximize the available technology will require both ongoing familiarity with technological advancements and a dash of cavalier creativity. Ecologists can not always be on top of such things and thus taking a molecular biologist friend out to lunch is an excellent idea, as you might have more in common than you can imagine.

Acknowledgements I would like to thank Jim Marden, Ilkka Hanski, Hans Ellegren, Tom Mitchell-Olds, Cris Vera, Heiko Vogel, Roger Butlin, Scott Edwards, Jessica Hellman, and Juan Galindo for the conversations, experience, and feedback on the ideas presented in this paper. Two anonymous reviewers also provided very insightful feedback and are thanked for their effort. Funding during the writing of this paper was supported by the Academy of Finland (grants numbers 38604 and 44887, Finnish Centre of Excellence Programme, 2006–2011) and grant EF-0412651 to J. H. Marden and I. Hanski from the US National Science Foundation.

References

- Abzhanov A, Kuo WP, Hartmann C, Grant BR, Grant PR et al (2006) The calmodulin pathway and evolution of elongated beak morphology in Darwin's finches. *Nature* 442:563–567. doi:10.1038/nature04843
- Agrafioti I, Stumpf MPH (2007) SNPSTR: a database of compound microsatellite-SNP markers. *Nucleic Acids Res* 35:D71–D75. doi:10.1093/nar/gkl806
- Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS (2007) SNP discovery via 454 transcriptome sequencing. *Plant J* 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. doi:10.1186/1471-2164-7-130
- Beldade P, McMillan WO, Papanicolaou A (2007) Butterfly genomics eclosing. *Heredity* 100:150–157. doi:10.1038/sj.hdy.6800934
- Bonaldo MF, Lennon G, Soares MB (1996) Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res* 6:791–806. doi:10.1101/gr.6.9.791
- Bouck A, Vision T (2007) The molecular ecologist's guide to expressed sequence tags. *Mol Ecol* 16:907–924. doi:10.1111/j.1365-294X.2006.03195.x
- Bourque G, Zdobnov EM, Bork P, Pevzner PA, Tesler G (2005) Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. *Genome Res* 15:98–110. doi:10.1101/gr.3002305
- Braby MF, Trueman JWH (2006) Evolution of larval host plant associations and adaptive radiation in pierid butterflies. *J Evol Biol* 19:1677–1690. doi:10.1111/j.1420-9101.2006.01109.x
- Brockman W, Alvarez P, Young S, Garber M, Giannoukos G et al (2008) Quality scores and SNP detection in sequencing-by-

- synthesis systems. *Genome Res* 18:763–770. doi:[10.1101/gr.070227.107](https://doi.org/10.1101/gr.070227.107)
- Carroll SB (2005) *Endless forms most beautiful: the new science of evo-devo*. W. W. Norton & Co, New York
- Chaisson M, Pevzner P, Tang H (2004) Fragment assembly with short reads. *Bioinformatics Oxf* 20:2067–2074. doi:[10.1093/bioinformatics/bth205](https://doi.org/10.1093/bioinformatics/bth205)
- Cheung F, Haas BJ, Goldberg SMD, May GD, Xiao YL et al (2006) Sequencing *Medicago truncatula* expressed sequenced tags using 454 life sciences technology. *BMC Genomics* 7:272. doi:[10.1186/1471-2164-7-272](https://doi.org/10.1186/1471-2164-7-272)
- Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R (2005) Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res* 15:1496–1502. doi:[10.1101/gr.4107905](https://doi.org/10.1101/gr.4107905)
- Ellegren H (2008) Sequencing goes 454 and takes large-scale genomics into the wild. *Mol Ecol* 17:1629–1635. doi:[10.1111/j.1365-294X.2008.03699.x](https://doi.org/10.1111/j.1365-294X.2008.03699.x)
- Ellegren H, Sheldon BC (2008) Genetic basis of fitness differences in natural populations. *Nat Rev Genet* 452:169–175. doi:[10.1038/nature06737](https://doi.org/10.1038/nature06737)
- Endler JA (1986) *Natural selection in the wild*. Princeton University Press, Princeton
- Ewing B, Green P (1998) Base-calling of automated sequencer traces using Phred II. Error probabilities. *PCR Methods Appl* 8:186–194
- Feder ME, Mitchell-Olds T (2003) Evolutionary and ecological functional genomics. *Nat Rev Genet* 4:651–657. doi:[10.1038/nrg1128](https://doi.org/10.1038/nrg1128)
- Feder ME, Watt WB (1992) Functional biology of adaptation. In: Crawford TJ, Hewitt GM (eds) *Genes in ecology*. Blackwell Scientific Publications, Oxford, pp 365–392
- Fulton TM, Van der Hoeven R, Eannetta NT, Taknsley SD (2002) Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. *Plant Cell* 14:1457–1467. doi:[10.1105/tpc.010479](https://doi.org/10.1105/tpc.010479)
- Gillespie JH (1991) *The causes of molecular evolution*. Oxford University Press, New York
- Goldberg SMD, Johnson J, Busam D, Feldblyum T, Ferreira S et al (2006) A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proc Natl Acad Sci USA* 103:11240–11245. doi:[10.1073/pnas.0604351103](https://doi.org/10.1073/pnas.0604351103)
- Haag CR, Saastamoinen M, Marden JH, Hanski I (2005) A candidate locus for variation in dispersal rate in a butterfly metapopulation. *Proc R Soc Biol Sci Ser B* 272:2449–2456. doi:[10.1098/rspb.2005.3235](https://doi.org/10.1098/rspb.2005.3235)
- Hanski I, Saccheri I (2006) Molecular-level variation affects population growth in a butterfly metapopulation. *PLoS Biol* 4:719–726. doi:[10.1371/journal.pbio.0040129](https://doi.org/10.1371/journal.pbio.0040129)
- Hanski I, Eralahti C, Kankare M, Ovaskainen O, Siren H (2004) Variation in migration propensity among individuals maintained by landscape structure. *Ecol Lett* 7:958–966. doi:[10.1111/j.1461-0248.2004.00654.x](https://doi.org/10.1111/j.1461-0248.2004.00654.x)
- Holt RA, Jones SJM (2008) The new paradigm of flow cell sequencing. *Genome Res* 18:839–846. doi:[10.1101/gr.073262.107](https://doi.org/10.1101/gr.073262.107)
- Hudson ME (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. *Mol Ecol Resour* 8:3–17. doi:[10.1111/j.1471-8286.2007.02019.x](https://doi.org/10.1111/j.1471-8286.2007.02019.x)
- Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM (2007) Accuracy and quality of massively-parallel DNA pyrosequencing. *Genome Biol* 8:R143
- Joron M, Papa R, Beltran M, Chamberlain N, Mavarez J et al (2006) A conserved supergene locus controls colour pattern diversity in *Heliconius* butterflies. *PLoS Biol* 4:1831–1840
- Kantety RV, La Rota M, Matthews DE, Sorrells ME (2002) Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum, and wheat. *Plant Mol Biol* 48:501–510. doi:[10.1023/A:1014875206165](https://doi.org/10.1023/A:1014875206165)
- Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F et al (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science* 318:420–426. doi:[10.1126/science.1149504](https://doi.org/10.1126/science.1149504)
- Lee J, Hever A, Willhite D, Zlotnik A, Hevezi P (2005) Effects of RNA degradation on gene expression analysis of human postmortem tissues. *FASEB J* 19:1356–1358. doi:[10.1096/fj.04-2591hyp](https://doi.org/10.1096/fj.04-2591hyp)
- Lévesque V, Fayad T, Ndiaye K, Nahé Diouf M, Lussier JG (2003) Size-selection of cDNA libraries for the cloning of cDNAs after suppression subtractive hybridization. *Biotechniques* 35:72–78
- Lewontin RC (1974) *The genetic basis of evolutionary change*. Columbia University Press, New York
- Long AD, Beldade P, Macdonald SJ (2007) Estimation of population heterozygosity and library construction-induced mutation rate from expressed sequence tag collections. *Genetics* 176:711–714. doi:[10.1534/genetics.106.063610](https://doi.org/10.1534/genetics.106.063610)
- Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nat Rev Genet* 4:981–994. doi:[10.1038/nrg1226](https://doi.org/10.1038/nrg1226)
- Lyons LA, Laughlin TF, Copeland NG, Jenkins NA (1997) Comparative anchor tagged sequences (CATS) for integrative mapping of mammalian genomes. *Nat Genet* 15:47–56
- Marden JH (2006) Quantitative and evolutionary biology of alternative splicing: how changing the mix of alternative transcripts affects phenotypic plasticity and reaction norms. *Heredity* 100:111–120
- Mardis ER (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet* 24:133–141
- Margulies M, Egholm M, Altman WE, Attiya S, Bader JS et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–380
- Meyer M, Stenzel U, Hofreiter M (2008) Parallel tagged sequencing on the 454 platform. *Nat Protocols* 3:267–278. doi:[10.1038/nprot.2007.520](https://doi.org/10.1038/nprot.2007.520)
- Mitchell-Olds T, Willis JH, Goldstein DB (2007) Which evolutionary processes influence natural genetic variation for phenotypic traits? *Nat Rev Genet* 8:845–856. doi:[10.1038/nrg2207](https://doi.org/10.1038/nrg2207)
- Moore MJ, Dhingra A, Soltis PS, Shaw R, Farmerie WG et al (2006) Rapid and accurate pyrosequencing of angiosperm plastid genomes. *BMC Plant Biol* 6:17. doi:[10.1186/1471-2229-6-17](https://doi.org/10.1186/1471-2229-6-17)
- Morin PA, Luikart G, Wayne RK, Sw group (2004) SNPs in ecology, evolution, and conservation. *Trends Ecol Evol* 19:208–216. doi:[10.1016/j.tree.2004.01.009](https://doi.org/10.1016/j.tree.2004.01.009)
- Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007) KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* 35:W182–W185. doi:[10.1093/nar/gkm321](https://doi.org/10.1093/nar/gkm321)
- Mountain JL, Knight A, Jobin M, Gignoux C, Miller A et al (2002) SNPSTRs: empirically derived, rapidly typed, autosomal haplotypes for inference of population history and mutational processes. *Genome Res* 12:1766–1772. doi:[10.1101/gr.238602](https://doi.org/10.1101/gr.238602)
- Nachman MW, Hoekstra HE, D’Agostino SL (2003) The genetic basis of adaptive melanism in pocket mice. *Proc Natl Acad Sci USA* 100:5268–5273. doi:[10.1073/pnas.0431157100](https://doi.org/10.1073/pnas.0431157100)
- Nair S, Williams JT, Brockman A, Paiphun L, Mayxay M et al (2003) A selective sweep driven by pyrimethamine treatment in southeast Asian malaria parasites. *Mol Biol Evol* 20:1526–1536. doi:[10.1093/molbev/msg162](https://doi.org/10.1093/molbev/msg162)
- Nielsen R (2001) Statistical tests of selective neutrality in the age of genomics. *Heredity* 86:641–647. doi:[10.1046/j.1365-2540.2001.00895.x](https://doi.org/10.1046/j.1365-2540.2001.00895.x)

- Noonan JP, Coop G, Kudaravalli S, Smith D, Krause J et al (2006) Sequencing and analysis of Neanderthal genomic DNA. *Science* 314:1113–1118. doi:10.1126/science.1131412
- Papanicolaou A, Gebauer-Jung S, Blaxter ML, McMillan DM, Jiggins CD (2008) Butterfly base: a platform for lepidopteran genomics. *Nucleic Acids Res* 36:D582–D587. doi:10.1093/nar/gkm853
- Paschall JE, Oleksiak MF, Van Wye JD, Roach JL, Whitehead JA et al (2004) Funny base: a systems level functional annotation of *Fundulus* ESTs for the analysis of gene expression. *BMC Genomics* 5:96. doi:10.1186/1471-2164-5-96
- Picoult-Newberg L, Ideker TE, Pohl MG, Taylor SL, Donaldson MA et al (1999) Mining SNPs from EST databases. *Genome Res* 9:167–174
- Pop M, Salzberg SL (2008) Bioinformatics challenges of new sequencing technology. *Trends Genet* 24:142–149
- Quinlan AR, Stewart DA, Strömberg MP, Marth GT (2008) Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nat Methods* 5:179–181. doi:10.1038/nmeth.1172
- Rosenblum EB, Novembre J (2007) Ascertainment bias in spatially structured populations: a case study in the eastern fence lizard. *J Hered* 98:331–336. doi:10.1093/jhered/esm031
- Saastamoinen M (2007) Heritability of dispersal rate and other life history traits in the Glanville fritillary butterfly. *Heredity* 100:39–46. doi:10.1038/sj.hdy.6801056
- Saastamoinen M, Hanski I (2008) Genotypic and environmental effects on flight activity and oviposition in the Glanville fritillary butterfly. *Am Nat* 171:701–712. doi:10.1086/587531
- Sarhan A (2006) Isolation and characterization of five microsatellite loci in the Glanville fritillary butterfly (*Melitaea cinxia*). *Mol Ecol Notes* 6:163–164. doi:10.1111/j.1471-8286.2006.01176.x
- Schmid KJ, Ramos-Onsins S, Ringys-Beckstein H, Weisshaar B, Mitchell-Olds T (2005a) A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. *Genetics* 169:1601–1615. doi:10.1534/genetics.104.033795
- Schmid M, Davison TS, Henz SR, Page UJ, Demar M et al (2005b) A gene expression map of *Arabidopsis thaliana* development. *Nat Genet* 37:501–506. doi:10.1038/ng1543
- Shiu S-H, Borevitz JO (2008) The next generation of microarray research: applications in evolutionary and ecological genomics. *Heredity* 100:141–149. doi:10.1038/sj.hdy.6800916
- Slate J (2005) Quantitative trait locus mapping in natural populations: progress, caveats and future directions. *Mol Ecol* 14:363–379. doi:10.1111/j.1365-294X.2004.02378.x
- Stein LD, Mungall C, Shu S, Caudy M, Mangone M et al (2002) The generic genome browser: a building block for a model organism system database. *Genome Res* 12:1599–1610. doi:10.1101/gr.403602
- Storz JF (2005) Using genomic scans of DNA polymorphism to infer adaptive population divergence. *Mol Ecol* 14:671–688. doi:10.1111/j.1365-294X.2005.02437.x
- Tishkoff SA, Verrelli BC (2003) Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annu Rev Genomics Hum Genet* 4:293–340. doi:10.1146/annurev.genom.4.070802.110226
- Torres TT, Metta M, Ottenwälder B, Schlötterer C (2008) Gene expression profiling by massively parallel sequencing. *Genome Res* 18:172–177. doi:10.1101/gr.6984908
- Van't Hof AE, Brakefield PM, Saccheri IJ, Zwaan BJ (2007) Evolutionary dynamics of multilocus microsatellite arrangements in the genome of the butterfly *Bicyclus anynana*, with implications for other Lepidoptera. *Heredity* 98:320–328. doi:10.1038/sj.hdy.6800944
- Vera C, Wheat CW, Marden JH, Hanski I (2007) Rapid transcriptome characterization for a non-model organism using massively parallel 454 pyrosequencing. *Mol Ecol* 17:1636–1647. doi:10.1111/j.1365-294X.2008.03666.x
- Vos P, Hogers R, Bleeker M, Reijans M, Vandelee T et al (1995) Aflp—a new technique for DNA fingerprinting. *Nucleic Acids Res* 23:4407–4414. doi:10.1093/nar/23.21.4407
- Wahlberg N, Wheat CW (2008) Genomic outposts serve the phylogenomic pioneers: designing novel nuclear markers for genomic DNA extractions of lepidoptera. *Syst Biol* 57:231–242. doi:10.1080/10635150802033006
- Watt WB (2003) Mechanistic studies of butterfly adaptations in ecology and evolution taking flight. In: Boggs CL, Watt WB, Ehrlich PR (eds) *Butterflies as model systems*. University of Chicago Press, Chicago, IL
- Weber APM, Weber KL, Carr K, Wilkerson C, Ohlrogge JB (2007) Sampling the arabidopsis transcriptome with massively parallel pyrosequencing. *Plant Physiol* 144:32–42. doi:10.1104/pp.107.096677
- Wheat CW, Watt WB, Pollock DD, Schulte PM (2006) From DNA to fitness differences: sequences and structures of adaptive variants of *Colias* phosphoglucose isomerase (PGI). *Mol Biol Evol* 23:499–512. doi:10.1093/molbev/msj062
- Wheat CW, Vogel H, Wittstock U, Braby MF, Underwood D et al (2007) The genetic basis of a coevolutionary key innovation. *Proc Natl Acad Sci USA* 104:20427–20431. doi:10.1073/pnas.0706229104
- Wheeler DA, Srinivasan M, Egholm M, Yufeng S, Chen L et al (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452:872–876. doi:10.1038/nature06884
- Wicker T, Schlagenhauf E, Graner A, Close TJ, Keller B et al (2006) 454 sequencing put to the test using the complex genome of barley. *BMC Genomics* 7:275. doi:10.1186/1471-2164-7-275
- Won Y-J, Sivasundar A, Wang Y, Hey J (2005) On the origin of Lake Malawi cichlid species: a population genetic analysis of divergence. *Proc Natl Acad Sci USA* 102:6581–6586. doi:10.1073/pnas.0502127102
- 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. doi:10.1186/1471-2164-8-5
- Wray GA (2007) The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet* 8:206–216. doi:10.1038/nrg2063
- Xia QY, Zhou ZY, Lu C, Cheng DJ, Dai FY et al (2004) A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science* 306:1937–1940. doi:10.1126/science.1102210
- 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. doi:10.1534/genetics.106.055541
- Zhang DX (2004) Lepidopteran microsatellite DNA: redundant but promising. *Trends Ecol Evol* 19:507–509. doi:10.1016/j.tree.2004.07.020
- Zhang Y, Austin CA (1997) Using rapid amplification of cDNA ends (RACE) to obtain full-length cDNAs. Humana Press, Clifton, UK
- Zhang DX, Hewitt GM (2003) Nuclear DNA analyses in genetic studies of populations: practice, problems, and prospects. *Mol Ecol* 12:563–584. doi:10.1046/j.1365-294X.2003.01773.x
- 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, Bogdanova EA, Shcheglov AS, Vagner LL, Khaspekov GL et al (2004) Simple cDNA normalization using kamchatka crab duplex-specific nuclease. *Nucleic Acids Res* 32:e37. doi:10.1093/nar/gnh031