MOLECULAR ECOLOGY

Molecular Ecology (2011)

NEWS AND VIEWS

MEETING REVIEW

# RAD in the realm of next-generation sequencing technologies

H. C. ROWE, S. RENAUT and A. GUGGISBERG
*Department of Botany, University of British Columbia, 3529-6270 University Blvd, Vancouver, BC V6T 1Z4, Canada*

**The first North American RAD Sequencing and Genomics Symposium, sponsored by Floragenex (http://www.floragenex.com/radmeeting/), took place in Portland, Oregon (USA) on 19 April 2011. This symposium was convened to promote and discuss the use of restriction-site-associated DNA (RAD) sequencing technologies. RAD sequencing is one of several strategies recently developed to increase the power of data generated via short-read sequencing technologies by reducing their complexity (Baird *et al.* 2008; Huang *et al.* 2009; Andolfatto *et al.* 2011; Elshire *et al.* 2011). RAD sequencing, as a form of genotyping by sequencing, has been effectively applied in genetic mapping and quantitative trait loci (QTL) analyses in a range of organisms including nonmodel, genetically highly heterogeneous organisms (Table 1; Baird *et al.* 2008; Baxter *et al.* 2011; Chutimanitsakun *et al.* 2011; Pfender *et al.* 2011). RAD sequencing has recently found applications in phylogeography (Emerson *et al.* 2010) and population genomics (Hohenlohe *et al.* 2010). Considering the diversity of talks presented during this meeting, more developments are to be expected in the very near future.**

*Keywords*: genotyping by sequencing, Illumina, restriction-site-associated DNA

*Received 20 May 2011; revision received 16 June 2011; accepted 20 June 2011*

RAD markers were initially used in conjunction with low-cost microarray genotyping resources (Miller *et al.* 2007), but the advent of massively parallel, next-generation sequencing technologies and concomitant drop in sequencing costs lead to the integration of short-read sequencing with RAD genotyping (Baird *et al.* 2008). In the case of RAD sequencing, ligation of sequencing adapters to restriction digested DNA prior to shearing focuses sequencing effort on tagged restriction sites, rather than randomly sequencing throughout the genome. This approach dramatically increases the coverage for a given sequenced site, increasing both the confidence in base identity and the likelihood that the same sites will be sequenced in multiple samples (Fig. 1). RAD sequencing provides an efficient method for the discovery of thousands of single nucleotide polymorphisms (SNPs). It was initially developed by Dr. Eric Johnson at the University of Oregon and later expanded into commercial services for plant (Floragenex) and 'other' (Biota Sciences) genome studies. This 'centre of origin' has led to the strong representation of fish and plants in studies employing RAD sequencing, as reflected in the organization of the symposium into 'Plant Genomics' and 'Marine and Conservation Genomics' sections, capped by a 'Technology and Bioinformatics' section. Yet as the use of RAD and similar techniques spreads among scientific communities, discourse may organize more naturally by data applications rather than organism studied. These applications might be divided into investigation of genome organization (linkage mapping, genome assembly, location of chromosome features such as centromeres and rearrangements) and population-level studies aimed at understanding the organization of intraspecific variation (identification of population structure, regions experiencing selection in particular environments, migration patterns and speciation).

Beyond a projected decrease in sequencing cost-per-sample owing to increased multiplexing capacities and sequence yield per lane, advances in RAD sequencing seem centered around the creative use of paired-end sequence reads (Fig. 1). One of the major limitations of current 'short-read' sequencing technologies is the difficulty of reliably assembling short sequence reads, to provide flanking sequence necessary for many SNP assays, especially in the absence of a reference genome. Both Eric Johnson (University of Oregon, Floragenex) and Rick Nipper (Floragenex) discussed improving local assembly of short nucleotide sequences using RAD sequence tags as a high-coverage anchors for sequence reads from the randomly sheared end of the same DNA molecule (Etter *et al.* 2011). Distances between RAD sequence tags and their paired ends can be increased via circularization techniques or partial digestion during library construction. These techniques then allow *de novo* assembly of contigs up to 5 kb in length. Bill Cresko (University of Oregon) also presented the development of transcriptome-based RAD (*eRAD*) to further reduce the complexity of transcriptomes.

RAD mapping approaches are particularly attractive for studying population diversity in organisms lacking reference genomes or possessing complex evolutionary histories. Several researchers discussed the use of RAD sequencing for salmonid conservation, genetic mapping or population-level studies. The salmonid fish lack a closely related reference genome, and this clade has experienced a

Correspondence: H. C. Rowe, Fax: (604) 822 6089;
E-mail: roweheat@gmail.com

**Table 1** Published studies using RAD sequencing

| Organism by application | References |
|---|---|
| Genetic mapping | |
| Three-spined stickleback (*Gasterosteus aculeatus*) | Baird *et al.* (2008) |
| Diamondback moth (*Plutella xylostella*) | Baxter *et al.* (2011) |
| Barley (*Hordeum vulgare*) | Chutimanitsakun *et al.* (2011) |
| Perennial Ryegrass (*Lolium perenne*) | Pfender *et al.* (2011) |
| *Neurospora crassa* | Baird *et al.* (2008) |
| Population genomics | |
| Three-spined stickleback (*Gasterosteus aculeatus*) | Hohenlohe *et al.* (2010) |
| Phylogeography | |
| Pitcher plant mosquito (*Wyeomyia smithii*) | Emerson *et al.* (2010) |
| Whole-genome sequencing | |
| *Escherichia coli* | Etter *et al.* (2011) |

recent whole-genome duplication (25–100 MYA) (Allendorf *et al.* 1975). Based on his work in pacific salmon, Jim Seeb (University of Washington) mentioned that these limitations render the identification of true SNPs through transcriptome sequencing problematic. In contrast, RAD sequencing enables relatively economical screening of a large number of markers (depending on the choice of restriction enzyme) and individuals (given the use of individual barcodes) with very high coverage. One demonstration on how RAD sequencing can circumvent some of these limitations comes from work in rainbow and cutthroat trout, where strict quality filtering drastically reduced the proportion of false-positive SNPs uncovered (Hohenlohe *et al.* 2011).

RAD data can be integrated with previously existing genetic resources, as in the case of barley, where mapping populations created from well-characterized doubled haploid stocks are associated with abundant phenotypic data. Pat Hayes (Oregon State University) discussed the use of RAD sequence data to develop a dense array of SNP-based markers for these lines. A broad range of previously identified markers, including phenotypic markers, isozymes, AFLPs and SSRs, were used as anchors and confirmation for the new genetic map based on RAD sequence SNPs. A total of 463 additional SNPs were identified, which not only increases resolution for previously identified quantitative trait loci, but also clarifies genomic architecture. Sequence-based markers clearly show that 'marker deserts'

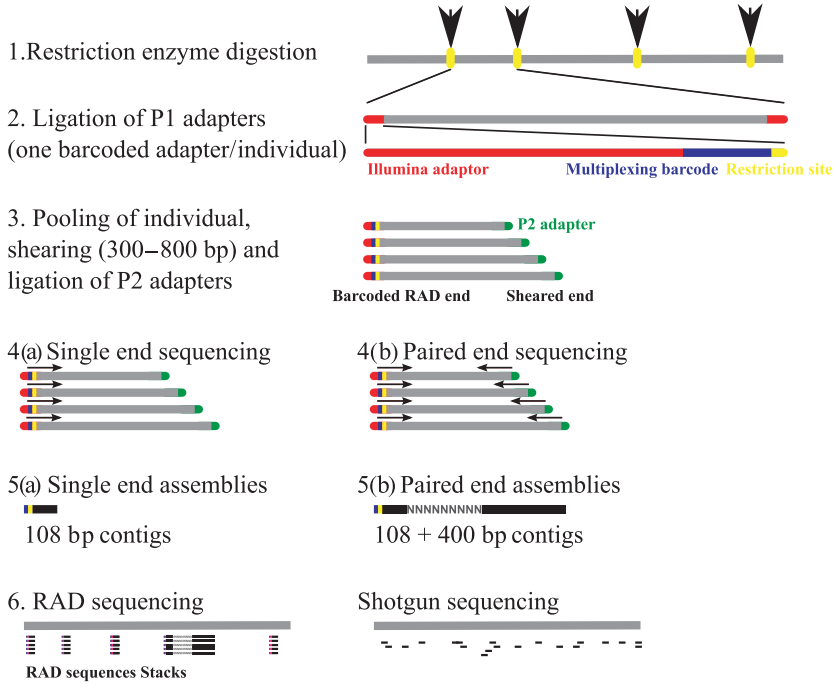


**Fig. 1** Overview of RAD sequencing. (1) DNA is digested with a restriction enzyme. (2) A modified adapter containing the Illumina P1 amplification and sequencing primer and a DNA barcode is ligated to the fragments. (3) Samples are pooled, sheared into 300- to 800-bp libraries (required for Illumina sequencing) and ligated to a second adapter P2. (4) Sequencing is performed either as (a) single end (one sequence of 36-108 bp per fragment) or (b) paired end (two sequences of 36-108 bp per fragment). (5a). Barcoded sequences are assembled into overlapping stacks. (5b) Given that restriction fragments are sheared randomly, paired-end sequencing allows the assembly of larger contigs on the sheared end of the fragment, whose size depends on the length of the Illumina sequences and the size of the sheared fragments isolated (see Etter *et al.* 2011 for more details and additional RAD sequencing strategies). (6a) Reduced representation of the genome through RAD sequencing, (b) as compared to shotgun sequencing. Partially redrawn from (Etter *et al.* 2011).

from previous maps were not caused by physical factors (such as rearrangements) that inhibit recombination, but instead result from a lack of polymorphism between the parental lines in these regions of the genome. As further examples of RAD sequence utility in clarifying existing genetic maps, RAD data were used in locating centromeres in salmonid genomes (Jim Seeb, University of Washington) and interspecific chromosomal rearrangements in annual sunflowers (Nolan Kane, University of British Columbia).

The majority of biologists can relate to the challenges posed by studying nonmodel organisms on a tight budget. As described by Kurt Lamour (University of Tennessee), RAD-sequence enabled genome-scale SNP genotyping of *Phytophthora capsaci* and has opened numerous avenues of investigation into the biology and spread of this relatively uncharacterized yet economically important plant pathogen. The success of this research highlights the economies of RAD sequencing approaches. Although the *P. capsaci* genome is relatively small (approximately 65 Mb), high rates of polymorphism (1 SNP/62 bp detected among seven isolates sequenced) make sequencing depth a priority for the discovery of reliable markers for genotyping. Use of economical genotyping technologies on these RAD-sequence-derived SNP markers allowed the assessment of previously unsuspected geographic differences in the population structure of this pathogen, with a high recombination rate detected in the central United States, suggesting the importance of the dormant, sexually produced oospore life stage for overwintering in harsh climates.

A primary goal of population genomics is identifying precise genomic locations subject to selective forces. Paul Hohenlohe (University of Oregon) presented recent population genomics work in stickleback that represents an important landmark for RAD sequencing (Baird *et al.* 2008; Hohenlohe *et al.* 2010). They have identified and genotyped tens of thousands of SNPs spread throughout the genome to confirm the association of candidate regions with adaptation to freshwater and, perhaps more importantly, identify additional regions that show parallel differentiation across independent freshwater populations. Using a similar approach, Bill Cresko and Paul Hohenlohe are using RAD sequencing to investigate stickleback adaptation following the extremely recent appearance and colonization (<60 years) of freshwater ponds in Alaska. Clearly, this body of work demonstrates some of the early breakthroughs in utilizing the capacity of RAD sequencing in identifying the effects of divergent selection on a genome-wide scale.

Efficiently processing the large amounts of data generated through massively parallel sequencing technologies is a subject of understandable interest to researchers. Julian Catchen from the University of Oregon presented the analytical tools provided in his open source software pipeline STACKS (http://creskolab.uoregon.edu/stacks/). STACKS was initially developed to construct genetic linkage maps from short sequence reads, but it can be used to identify SNPs in natural populations for subsequent population genomic analyses (Hohenlohe *et al.* 2010) or phylogeographic investigations (Emerson *et al.* 2010). The structure of STACKS facilitates data mining and data correction, as its output is meant to be viewed in the open source database MySQL (http://www.mysql.com/). A typical workflow for *de novo* RAD analyses would include the successive use of the program *process_radtags* to filter out bad-quality reads using a sliding-window approach, *ustacks* to align the reads into exactly matching stacks and detect SNPs by comparing those stacks in a likelihood framework, *cstacks* to build a catalogue of consensus sequences from the $F_0$ parents, *sstacks* to match each $F_1$ progeny against the catalogue and the Perl script *genotypes.pl* to export the genotypes to an external mapping program (e.g. R/qtl; Broman *et al.* 2003).

One of the strengths of RAD sequencing for genome-scale research projects is its generality, as, beyond the practicalities of experimental design and nucleic acid extraction, many approaches are equally applicable to organisms ranging from barley to salmon. In addition, RAD sequencing allows smaller research groups, or groups studying organisms that do not yet possess a reference genome, to conduct 'genomewide studies'. This revolution is perhaps similar to the one that took place nearly 20 years ago with the development of AFLP markers (Vos *et al.* 1995). In the present case, however, markers emerge from anonymity thanks to the genotyping-by-sequencing approach.

## References


Allendorf FW, Utter FM, May BP (1975) Gene duplication within the family Salmonidae: II. Detection and determination of the genetic control of duplicate loci through inheritance studies and the examination of populations. In:*Isozymes* (ed Marken CL). pp. 415–432, Academic Press, London.

Andolfatto P, Davison D, Erezyilmaz D *et al.* (2011) Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Research*, **21**, 610–617.

Baird NA, Etter PD, Atwood TS *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**, e3376.

Baxter SW, Davey JW, Johnston JS *et al.* (2011) Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. *PLoS ONE*, **6**, e19315.

Broman KW, Wu H, Sen S, Churchill GA (2003) R/qtl: QTL mapping in experimental crosses. *Bioinformatics Applications Note*, **19**, 889–890.

Chutimanitsakun Y, Nipper RW, Cuesta-Marcos A *et al.* (2011) Construction and application for QTL analysis of a restriction site associated DNA (RAD) linkage map in barley. *BMC Genomics*, **12**, 4. doi:10.1186/1471-2164-1112-1184.

Elshire RJ, Glaubitz JC, Sun Q *et al.* (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, **6**, e19379.

Emerson KJ, Merz CR, Catchen JM *et al.* (2010) Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Science, USA*, **107**, 16196–16200.

Etter PD, Preston JL, Bassham S, Cresko WA, Johnson EA (2011) Local *de novo* assembly of RAD paired-end contigs using short sequencing reads. *PLoS ONE*, **6**, e18561.

Hohenlohe PA, Bassham S, Etter PD *et al.* (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLos Genetics*, **6**, e1000862.

Hohenlohe PA, Amish SJ, Catchen JM, Allendorf FW, Luikart G (2011) Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources*, **11**, 117–122.

Huang X, Feng Q, Qian Q *et al*. (2009) High-throughput genotyping by whole-genome resequencing. *Genome Research*, **19**, 1068–1076.

Miller MR, Dunham JP, Amores JP, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated (RAD) markers. *Genome Research*, **17**, 240–248.

Pfender WF, Saha MC, Johnson EA, Slabaugh MB (2011) Mapping with RAD (restriction-site associated DNA) markers to rapidly identify QTL for stem rust resistance in *Lolium perenne*. *Theoretical and Applied Genetics*, **122**, 1467–1480.

Vos P, Hogers R, Bleeker M *et al*. (1995) AFLP: a new technique for DNA fingerprinting. *Nucleic Acid Research*, **23**, 4407–4414.

---

H.C.R. studies how environmental factors influence genetic diversity, with a current focus on secondary chemistry of hybrid sunflowers. S.R. focuses on using next generation sequencing data to decipher the genetic basis of species divergence both in salmonids and sunflowers. A.G. is using Canada thistle as a model system to investigate the genetic changes underlying the evolution of invasiveness in weedy plants.

---