New approaches for local *de novo* assembly and haplotype discovery using short sequencing reads

Paul Etter

Institute of Molecular Biology, University of Oregon
Restriction-site Associated DNA (RAD) markers

Restriction sites in genome

Mike Miller, Nate Baird, Tressa Atwood
Restriction-site Associated DNA (RAD) markers

P1 adapter

Forward flow cell priming site

Single-end sequencing primer

Barcoded for multiplex identification
Restriction-site Associated DNA (RAD) markers

Restriction sites in genome

P1 adapter
Forward flow cell priming site
Single-end sequencing primer
Barcoded for multiplex identification

Divergent P2 adapter
Reverse flow cell priming site

Mike Miller, Nate Baird, Tressa Atwood
Restriction-site Associated DNA (RAD) markers

P1 adapter

Forward flow cell priming site

Single-end sequencing primer

Barcoded for multiplex identification

Individual 1

Individual 2

Divergent P2 adapter

Restriction sites in genome

Reverse flow cell priming site

Mike Miller, Nate Baird, Tressa Atwood
• Recapitulated previous linkage analyses in stickleback, mapped induced mutation in Neurospora

• Demonstrated several advantages of RAD markers
RAD markers

- Reliably sample a subset of the genome at different densities by enzyme choice

- Multiplexed sample prep increases throughput & saves labor

- Allow rapid genetic marker discovery & simultaneous genotyping in multiple individuals, even without a reference genome

- Reduced sequence space allows rapid & accurate alignment, plus high depth of coverage for SNP & indel calls
RAD today

• Subsequent papers confirm it’s utility

• population genomics, genetic linkage maps, de novo assembly & haplotype sequencing
Pondering directions for RAD technology

Restriction sites in genome

RAD tag sequence read

Sheared-end reads
Pondering directions for RAD technology

Restriction sites in genome

RAD tag sequence read

Sheared-end reads

Variable length RAD fragments isolated
Local assembly from genome-wide data with paired-end RAD

Restriction sites in genome

RAD tag sequence read

Sheared-end reads

Variable length RAD fragments isolated

P1 adapter
Forward flow cell priming site
Single-end sequencing primer
Barcoded for multiplex identification

Divergent P2 adapter
Reverse flow cell priming site
Paired-end sequencing primer

Contigs assembled from the sheared-end reads for each RAD tag
A test case

• Two individuals from a polymorphic stickleback population

• Barcoded libraries prepared & sequenced to high depth of coverage
RAD paired-end (RAD-PE) contigs in stickleback
RAD paired-end (RAD-PE) contigs in stickleback
SNP & indel detection

Low lateral plate individual

High lateral plate individual
SNP & indel detection

Low lateral plate individual

High lateral plate individual
SNP & indel detection
Haplotype analysis
Haplotype analysis

• Polymorphisms that occur on the same chromosome that tend to be inherited together

• Important implications for human medicine, understanding gene expression and population structure

• Genetic variation can only truly be understood from complete phase information across the genome
Haplotype analysis

SNP in RAD site

SNPs in contigs
New uses for a budding technique
New uses for a budding technique

• Overlapping contigs for *de novo* assembly using partial-digests with frequent cutters
Overlapping contigs with partial-digest RAD-PE

Restriction sites in genome

Full digest

Partial digest
Overlapping contigs with partial-digest RAD-PE (E. coli)
Overlapping contigs with partial-digest RAD-PE (*E. coli*)

• Cutting a trapezoidal slice during final gel extraction improves coverage across the full length of contigs

• Choosing only the longest contig built from each RAD site increases N50 to 729 bp
How about even longer contigs?
Long-insert RAD-PE contigs with a circularization step

1-6 kilobase RAD fragments isolated

Circularize & shear

Circularize & amplify
Up to 5 kb contigs with Long-insert RAD-PE (E. coli)
Up to 5 kb contigs with Long-insert RAD-PE (E. coli)
RAD paired-end sequencing

- For genotyping
  - Increased sequence space for marker detection
  - Haplotype discovery & analysis
  - Coding sequence discovery & analysis
  - Primer design for high-throughput genotyping platforms
- *de novo* assembly
  - Local assembly
  - Error-free PacBio/3rd-generation read lengths from existing short read technology
What’s new / who cares about haplotypes anyway?
What’s new / who cares about haplotypes anyway?

- One SNP, two SNP, both SNPs…disease

- Expression QTLs & deleterious coding variants impact one another & may be the more highly expressed haplotype (Lappalainen et al.)

- Population structure inference could be improved (Gattepaille & Jakobsson)

- Haplotypic diversity confounds assembly
Dilution haplotype sequencing
Dilution haplotype sequencing

Genomic DNA
Dilution haplotype sequencing

Genomic DNA

Shear

2 kb fragments
Dilution haplotype sequencing

Genomic DNA

Shear

2 kb fragments

Dilute to ~1/3 haploid genome-equivalent aliquots
Dilution haplotype sequencing

Genomic DNA → Shear → 2 kb fragments

Dilute to ~1/3 haploid genome-equivalent aliquots
Dilution haplotype sequencing

Genomic DNA

Shear

2 kb fragments

Dilute to ~1/3 haploid genome-equivalent aliquots

Amplify, shear again & create barcoded Illumina ‘shotgun’ libraries
Dilution haplotype sequencing

Genomic DNA → Shear → 2 kb fragments

Dilute to ~1/3 haploid genome-equivalent aliquots

Amplify, shear again & create barcoded Illumina ‘shotgun’ libraries

Sequence and align reads
Dilution haplotype sequencing

A test case

• 24 dilutions, 2-3 kb ‘Adapted’ stickleback DNA fragments from one individual (Bear Paw lake), amplified separately ~ 500 femtograms each

• Barcoded Illumina libraries were created from each & samples were combined after adapter ligation

• Sequenced on half HiSeq lane, ~85M barcoded reads

Bill Cresko, Mark Currey
Dilution haplotype sequencing
Dilution haplotype sequencing

SNPs in coding sequences as well as up and downstream regions
Dilution haplotype sequencing

chrIV 13013000 TGACTACATCTTCTCAAAATTATAATATCTTCCTCAGACACAACAAACCGCCAAAACCTCT
chrIV 13013000 SNPs AC C T T A
chrIV 13013000 SNPs
chrIV 13013000 SNPs
chrIV 13013000 ACGTAT
chrIV 13013000 ACTAAT ..................ac.c:::T::A:...........................................
chrIV 13013000 ACTCCT ...............................T:..............................................
chrIV 13013000 ATGCTT
chrIV 13013000 ATTATT
chrIV 13013000 ATTCAT
chrIV 13013000 CCGGTT
chrIV 13013000 CCTGCT .................................
chrIV 13013000 CGAACT ..........................................
chrIV 13013000 CTTGCT
chrIV 13013000 CTTCGT
chrIV 13013000 GAAACT
chrIV 13013000 GCCATT ..........................:T:..............................................
chrIV 13013000 GCCCGT
chrIV 13013000 GCCTAT ..........................................
chrIV 13013000 GTAGCT
chrIV 13013000 GTGCCT
chrIV 13013000 GTGGAT
chrIV 13013000 TCTGAT ....................:AC:C:::T::A:...........................................
chrIV 13013000 TCTTCT
chrIV 13013000 TGACAT
chrIV 13013000 TGCGGT ..........................................
chrIV 13013000 TTGGAT
chrIV 13013000 TTTAGT .........................ac.c...t..a......................................

Haplotype 1 TGACTACATCTTCTCAAAATTATAATATACTCTTTACACAGACACAACAAACCGCCAAAACCTCT
Haplotype 2 TGACTACATCTTCTCAAAATTATAATATCTTTTTTTCACTCAGACACAACAAACCGCCAAAACCTCT
Dilution haplotype sequencing Summary

• The approach works

• Need to improve coverage across the genome & within 2 kilobase blocks
Potential uses & future directions

• Correlation with allele-specific gene expression

• ‘Finishing’ genome assemblies

• Potential for individual, cheap & rapid haplotyping of individuals with a minimum of resources
Thanks

- Johnson lab members - Doug Turnbull, Nick Kamps-Hughes, Jessica Preston, Jason Carriere & Eric

- Collaborators - Cresko lab, Mike Miller, Postlethwait lab, Floragenex

- UO Genomics Facility – Doug Turnbull & Nicholas Stiffler

- Funding – NIH, NHGRI & NIGMS

- Illumina