UNIVERSITÄT DUISBURG ESSEN

Open-Minded

Universitätsklinikum Essen
Humangenetik / Genominformatik

Simulation of Double Digest Restriction Site Associated DNA Sequencing Data

Henning Timm[1], Hannah Weigand[2], Martina Weiss[2], Florian Leese[2] & Sven Rahmann[1]

1: Genome Informatics, Institute of Human Genetics, University of Duisburg-Essen, University Hospital Essen, Essen, Germany
2: Aquatic Ecosystem Research, University of Duisburg-Essen, Essen, Germany
Contact: henning.timm@tu-dortmund.de

Motivation

The availability of low cost sequencing technologies allows the evaluation of thousands of genetic markers spread across the genomes of non-model organisms, across populations.
Due to the specific structure of ddRADseq data, testing analysis tools is not a trivial task. When using real data, the ground truth, including the genotypes of the individuals, the locus sequences and the number of loci, is not known.
RAGE, the ddRAD Data Generator, solves this problem by simulating ddRAD reads using a locus based model and providing a detailed, well annotated ground truth.

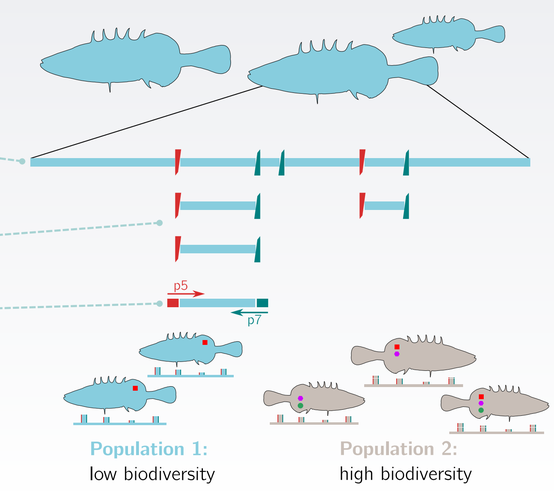
ddRAD Sequencing

Reduced representation sequencing technique, used for inter- and intra-population genotyping studies.

- Cut DNA into fragments using restriction enzymes:
  Rare cutter Frequent cutter
- Filter fragments by length and (restr. site) structure
- Add primers and auxiliary sequences
- Paired-end sequencing (Illumina)

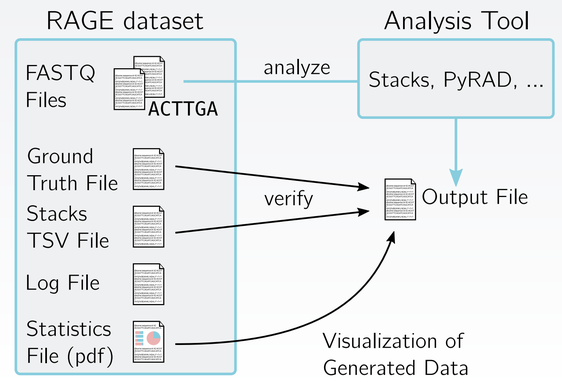
A pair of restriction sites that produces reads is called a **locus**.

Loci appear at the same position in different individuals of the same species.



Why simulate ddRAD data?

Provide an easily verifiable ground truth that can be used to evaluate analysis pipelines.



Simulation Workflow

Simulation process to create a RAGE dataset:

Preprocessing

1 Pick Individuals
- Specify auxiliary sequences

2 Create Template Reads
- Create sequence for each locus
- Add (still ambiguous) DBR
- Add enzyme residues

Read Simulation

3 Pick Individual Event Type
- Specify read behavior at locus (`Dropout`, `Common` or `SNP`)
- Specify subtype (NA, zygosity)
- Create reads for picked type

4 Finalize Reads
- Set unambiguous DBR sequence
- Join fragments of reads

5 PCR copies
- Add copies with duplicate DBRs

Postprocessing

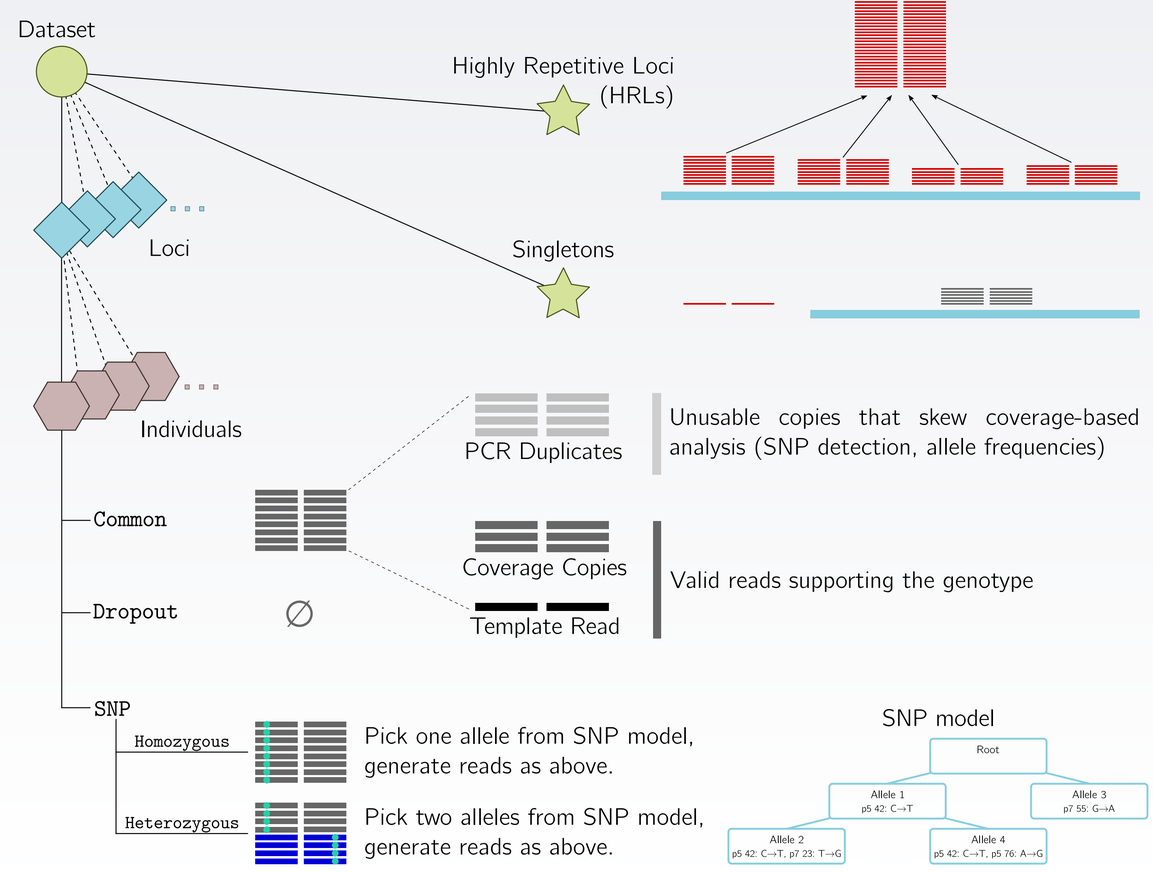
6 Singletons
- Add reads without related locus

7 HRL Reads
- Add reads from highly repetitive loci

8 Sequencing Errors
- Change bases with probability p=0.01

Each run generates:
- FASTQ files with simulated reads
- TSV files containing the simulated effects
- Log files and (visual) dataset statistics

Dataset Structure



Individual Event Types

`Common` ~90%
Reads without coverage deviation and mutations.

`Dropout` ~5%
No reads for an individual at a locus due to sequencing errors or a p5 null allele.

`SNP` ~5%
One or more point mutations. The coverage is equally distributed between both alleles.

Individual Event Subtypes

Zygosity
- `Homozygous` - `Heterozygous`

`SNP` type events can either be homozygous or heterozygous. One or two alleles are chosen from the SNP model.

Null Allele
- `Mutation` - `Incomplete Digestion`

Different p7 sequence for a `common` or `SNP` event. Either due to mutation of the p7 restriction site or an undigested p7 restriction site.

Coverage in ddRAD Data

Coverage := number of reads per locus
The coverage varies with sequencing technology, library preparation and biological effects. For an individual $i \in \mathcal{I}$ at locus $\ell \in \mathcal{L}$ the coverage is denoted as:

$$\mathrm{cov}_{i,\ell} \left| \begin{array}{l} \mathrm{cov}_{i,\ell}(a_1) \\ \mathrm{cov}_{i,\ell}(a_2) \end{array} \right.$$

The observed coverage in rad data mainly depends on the targeted sequencing depth d_s.
In real data the coverage seldomly reaches d_s.
Hence, the coverage is simulated as a function of d_s using a probabilistic process.

Distribution of coverage values in a real ddRAD dataset with 35 million reads. PCR duplicates have not been removed.

All loci, including singletons.

Loci of size >= 3.

Coverage Model

To account for the specific behavior of different groups of reads, three different models are used:

Coverage for singletons

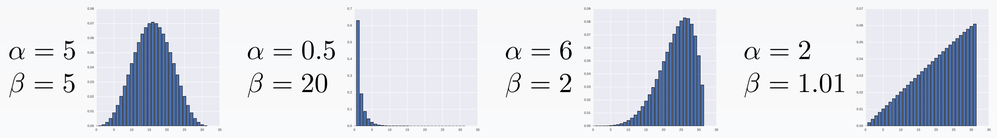
Singletons, by definition, have a coverage of 1. The amount of singletons varies with the size of the dataset and the quality of library preparation. Hence, a fraction of the expected number of reads is used to simulate singleton reads.

Coverage for valid reads

The distribution of coverage values varies with different datasets due to library preparation, biological effects, and the sequencing process. Hence, an adaptable function is needed for sampling coverage values. This is solved using a beta-binomial distribution (BBD). The BBD has three parameters:

$\alpha > 0$	Shape parameter controlling the left tailing, higher value → more left-skew
$\beta > 0$	Shape parameter controlling the right tailing, higher value → more right-skew
$n \in \mathbb{N}_0$	Number of trials, maximum number of events

Using the shape parameters, the BBD can be fitted to many different observed coverage patterns:



In order to make d_s the expected coverage value, the mean of the BBD has to be moved. This is done by chosing the n parameter depending on α and β:

$$\mathbb{E}(X) = \frac{n\alpha}{\alpha+\beta} \qquad X \sim \mathrm{BBD}(\alpha, \beta, n) \qquad n = \left\lfloor \frac{d_\mathrm{s} \cdot (\alpha+\beta)}{\alpha} \right\rceil$$

Coverage for HRLs

The coverage of HRLs can reach values beyond 1000. To simulate this, a distribution with high variance is needed. Finding a distribution that models the observed values well is still a topic of research. Both a Poisson distribution and a discrete uniform distribution are currently being evaluated.

Simulated Coverage Distribution

Example of combined simulated coverage using α=6, β=2.5, n=42, d_s=30, a uniform model for HRLs and no PCR duplicates:

Future Work

Adding PCR duplicates will smooth out the distinct shapes and approach the observed coverage distributions.

Choosing BDD parameters can be automated by (algorithmically) fitting the distribution to observed coverage distributions in sample datasets.

A more accurate model for the simulation of HRL coverages still needs to be found.

References

J. Catchen, P. Hohenlohe, S. Bassham, A. Amores, W. Cresko (2013). Stacks: an analysis tool set for population genomics. Molecular Ecology, **22**(11), 3124–3140.
D. Eaton (2014). PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. Bioinformatics, **30**(13), 1844–1849.