Simulating ddRADseq Reads with RAGE

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

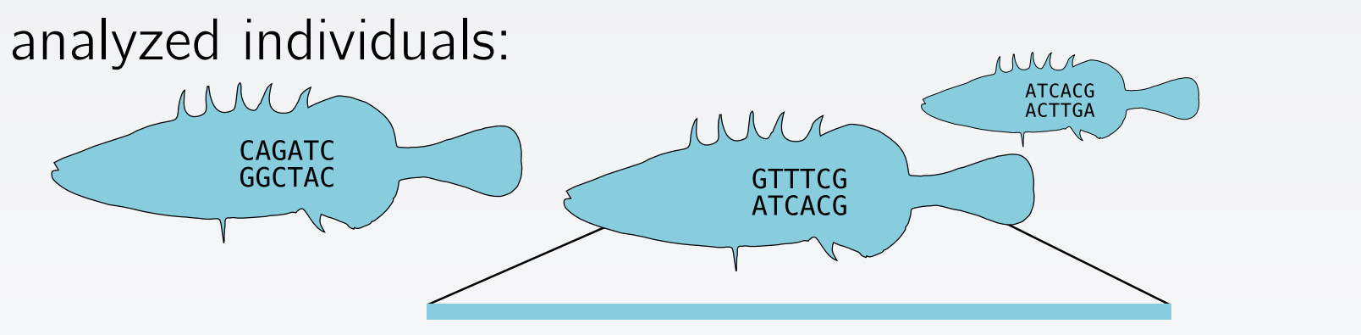
1: Genome Informatics, Institute of Human Genetics, University of Duisburg-Essen, University Hospital Essen, Essen, Germany
2: Aquatic Ecosystem Research, University of Duisburg-Essen, Essen, Germany
Contact: henning.timm@tu-dortmund.de

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 and well annotated ground truth.

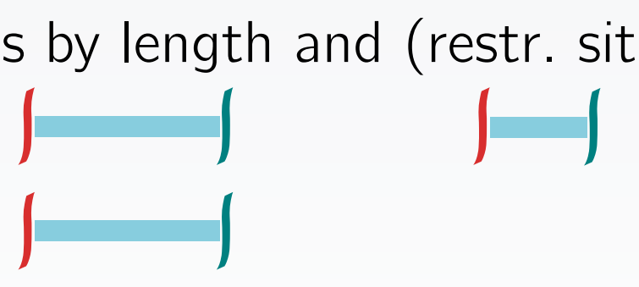
ddRAD Sequencing

Reduced representation sequencing technique, applied to reduce costs and maximize number of analyzed individuals:

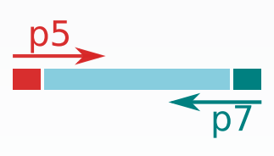


- Cut DNA into fragments using restriction enzymes:

- Filter fragments by length and (restr. site) structure



- Add primers and auxiliary sequences
- Paired-end Sequencing (Illumina)

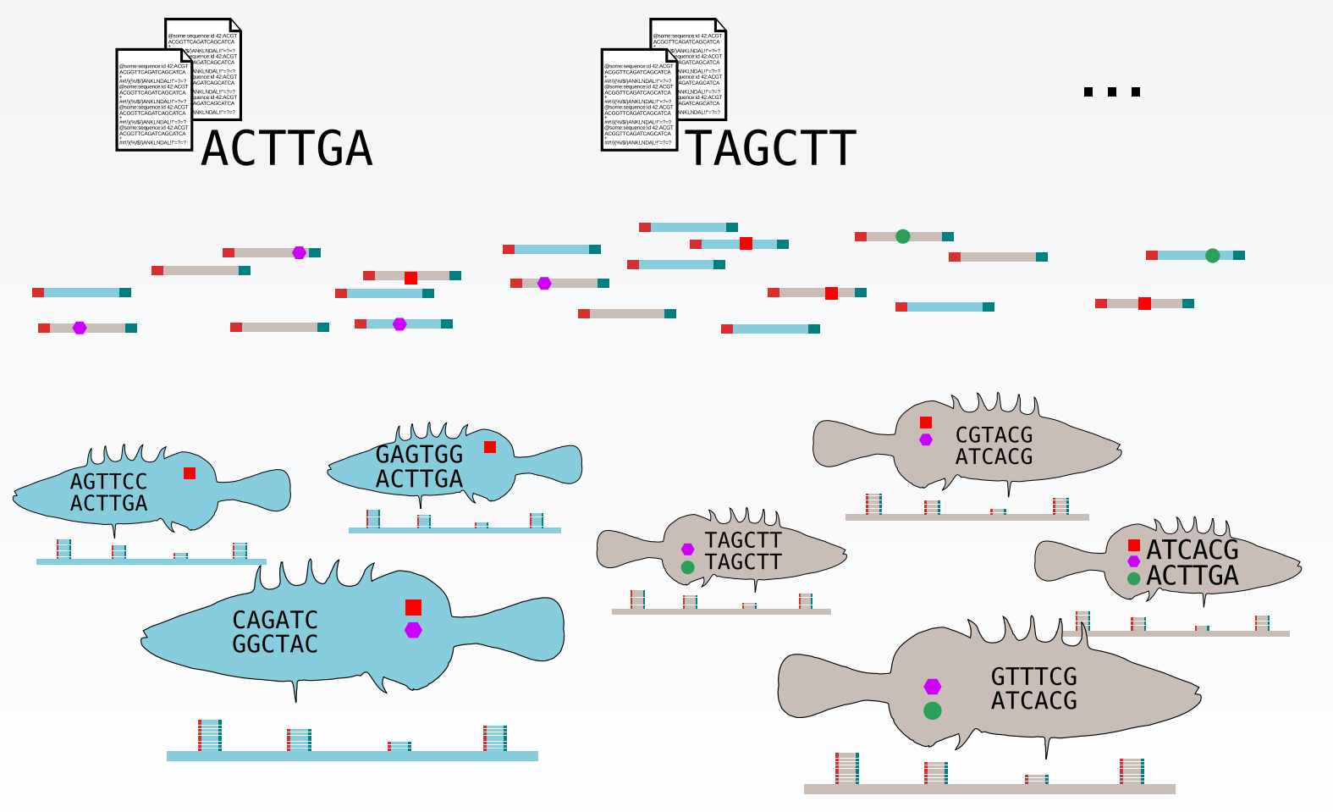


A pair of restriction sites that produces reads is called a **locus**.

ddRAD Analysis

Given: Reads from several individuals as FASTQ files.

Goals:
- Call SNVs at loci
- Identify populations
- Infer genetic diversity
- Inter-pop. analysis



Tools: Stacks [Catchen et al. (2013)], PyRAD [Eaton (2014)].

Other Work

While simulation tools for ddRAD data exist, they do not provide a general verifiable ground truth:

ddRADseqtools [Mora-Márquez et al. (2016)]

Simulates ddRAD pipeline to optimize pipeline parameters. Generates reads from a reference genome, but does not create a ground truth.

simrad [Lepais et al. (2014)]

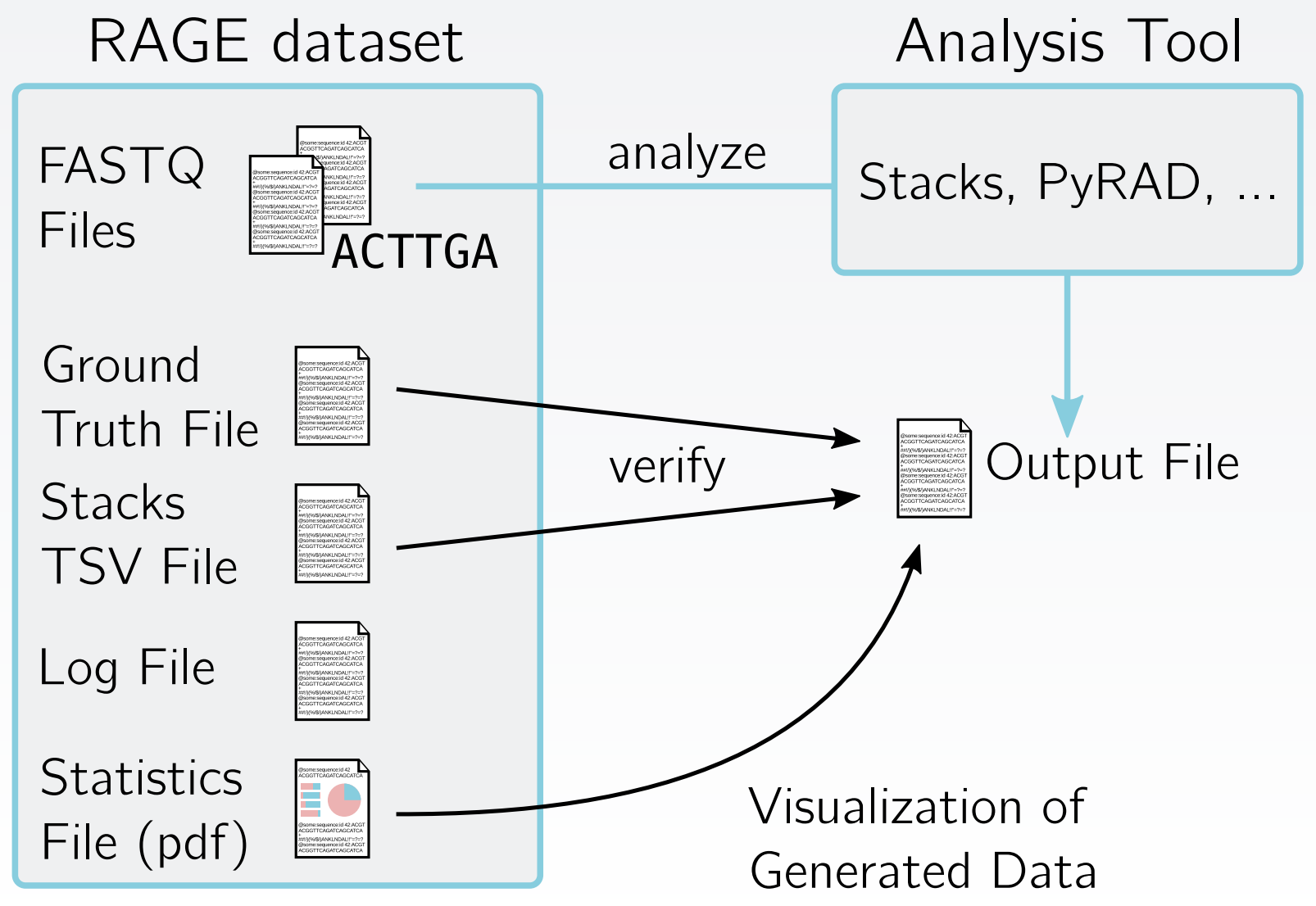
R library simulating the digestion process to predict the number of created loci and pick optimal parameters for a RAD experiment.

simrrls [Eaton (2014)]

Used to test PyRAD. Simulates ddRAD reads, but does not create a detailed ground truth.

Main Goal

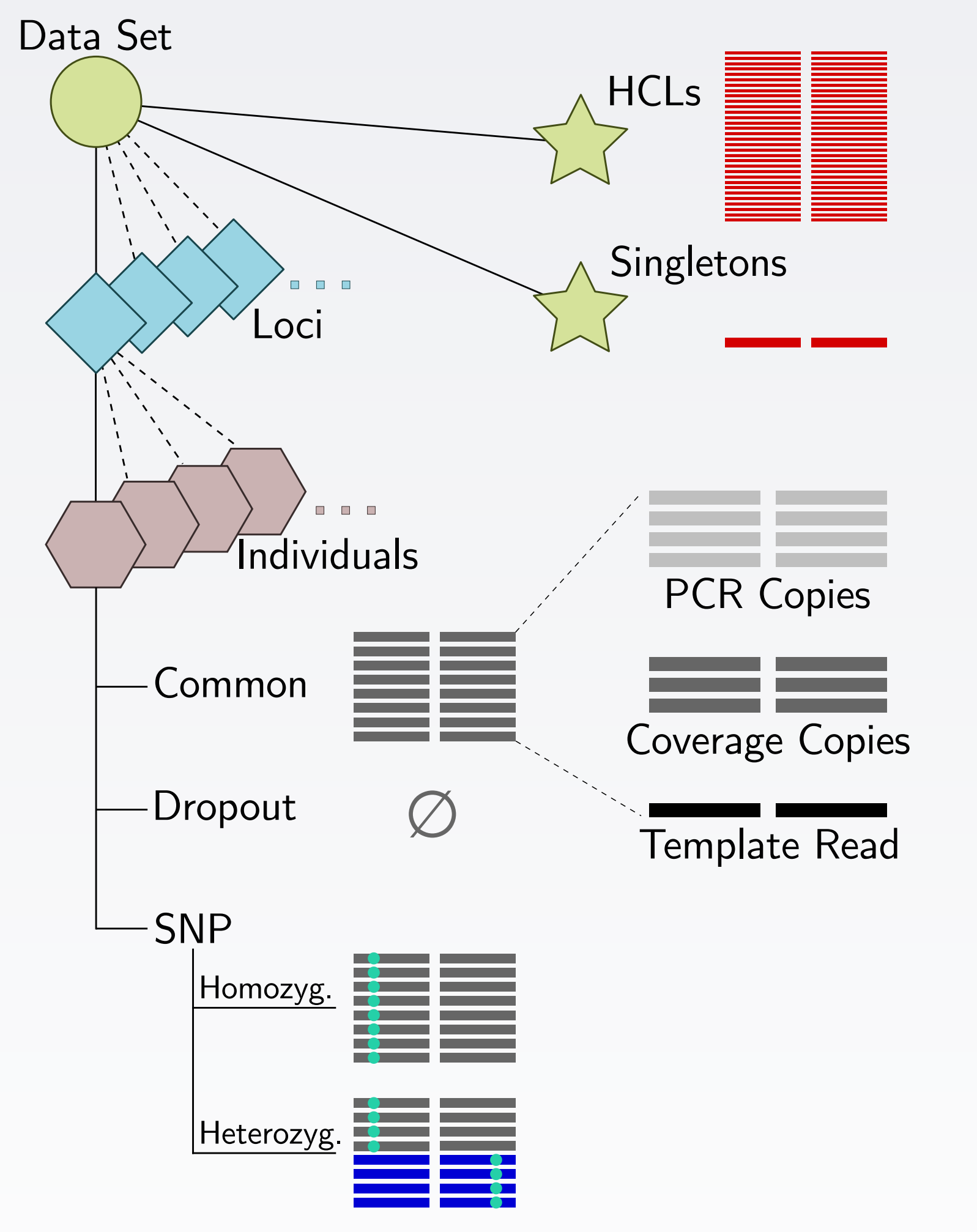
Provide an easily verifiable ground truth that can be integrated into an analysis pipeline.



Read Simulation Workflow

RAGE uses a locus-centric model to simulate ddRAD reads and adds ddRAD specific effects and errors to accurately mimic the structure of ddRAD data sets. The structure of a complete data set is illustrated on the right. The simulation steps, which are executed once **per data set**, **per locus** or **per individual**, are as follows:

1. Pick Individuals
   - Specify barcodes and inserts
2. Create Template Reads
   - Create sequence for each locus
   - Add (still ambiguous) DBR
   - Add enzyme residues
3. Pick Event Type
   - Specify read behavior at locus (Dropout, Common or SNP)
   - With or without Null Allele
   - Create proto-reads for picked type
4. Finalize Reads
   - Set unambiguous DBR sequence
   - Join fragments of proto-reads
5. PCR copies
   - Add copies with duplicate DBRs
6. Singletons
   - Add reads without related locus
7. HCL Reads
   - Add loci with high coverage
8. Sequencing Errors
   - Change bases with probability p

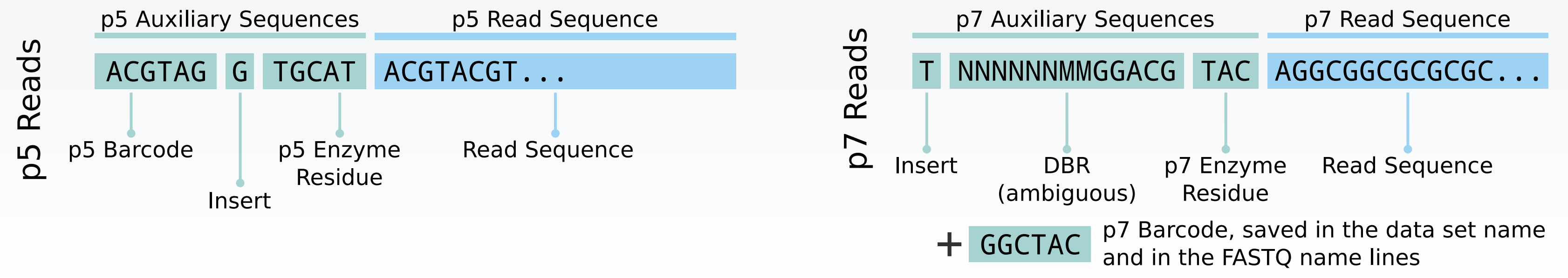


The finished data set contains reads with intrinsic (biological) and extrinsic (techical) features that have to be detected by analysis tools. While the reads created at loci carry useful information, HCL reads, singletons, and PCR duplicates are noise that has to be removed prior to analysis.

Read Structure

Each ddRAD read consists of auxiliary sequences, added by the sequencing process, and a read sequence, the individual's DNA sequence.

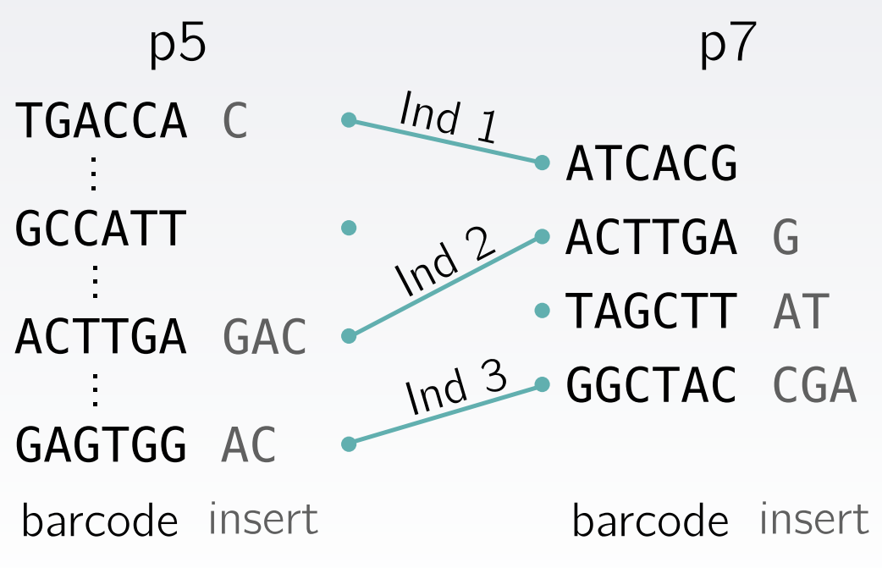
Due to the paired-end sequencing approach used by ddRADseq, p5 (forward) and p7 (reverse) reads are generated:



p7 Barcode, saved in the data set name and in the FASTQ name lines

Individual Selection (Step 1)

Multiple individuals in a sample are multiplexed using barcode pairs. A barcode file contains all valid barcode combinations and the associated individuals, as well as insert sequences (used to improve sequencing quality). RAGE uses the Illumina TrueSeq LT-Kit barcodes as default.
As samples are split up by p7 barcodes in the sequencing pipeline, all individuals in a RAGE data set share the same p7 barcode. Individuals are uniformly randomly selected from the barcode file.



Sequence Generation (Step 2)

For each locus, a template read sequence is generated, which is shared by all individuals at the locus. This sequence is guaranteed to not contain any restriction site for the enyzmes used in the ddRAD process. Per default, all bases are chosen uniformly, but the simulation can be adjusted to use different levels of GC content.

Event Types (Step 3)

The behavior of loci can be classified into 3 categories, which are chosen based on their abundance in real data. For each individual at each locus the event type is one of:

Common ~90%
Reads without any modifications are added, reaching the target coverage.

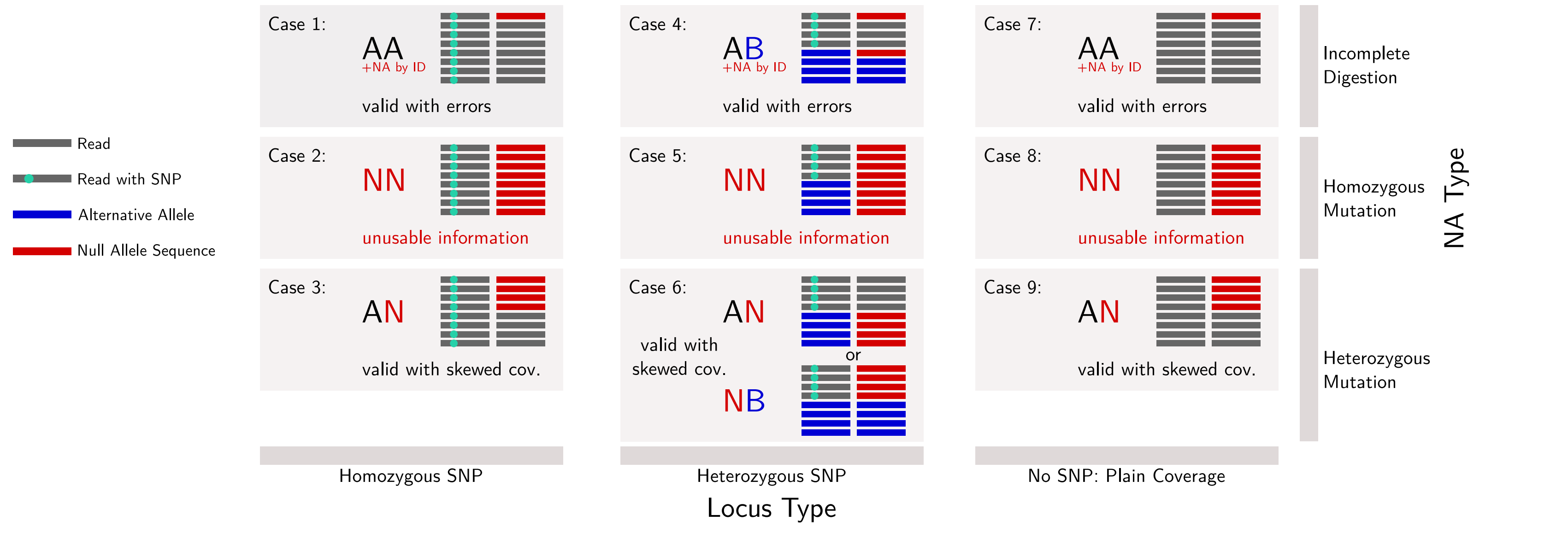
Dropout ~5%
No reads for the individual at this locus due to sequencing errors or a p5 null allele.

SNP ~5%
A homozygous or heterozygous SNP. The coverage is equally distributed between both alleles.

Common and SNP events can be affected by a null allele (NA), which modifies the p7 sequence. This can alter the coverage detected by analysis tools and break up loci, thereby hindering the analysis. There are two types of NAs, classified by their origin:

Incomplete Digestion: A restriction site has not been digested, altering the fragment boundaries and completely changing the p7 sequence. The p7 sequences of affected reads do not match the rest of the reads, but not all reads of an individual are affected.

Mutation: A mutation has altered the restriction site. The individual has a p7 sequence that consistently differs from others at the locus. If the mutation is heterozygous, only half the reads are affected.



PCR Copies (Step 5)

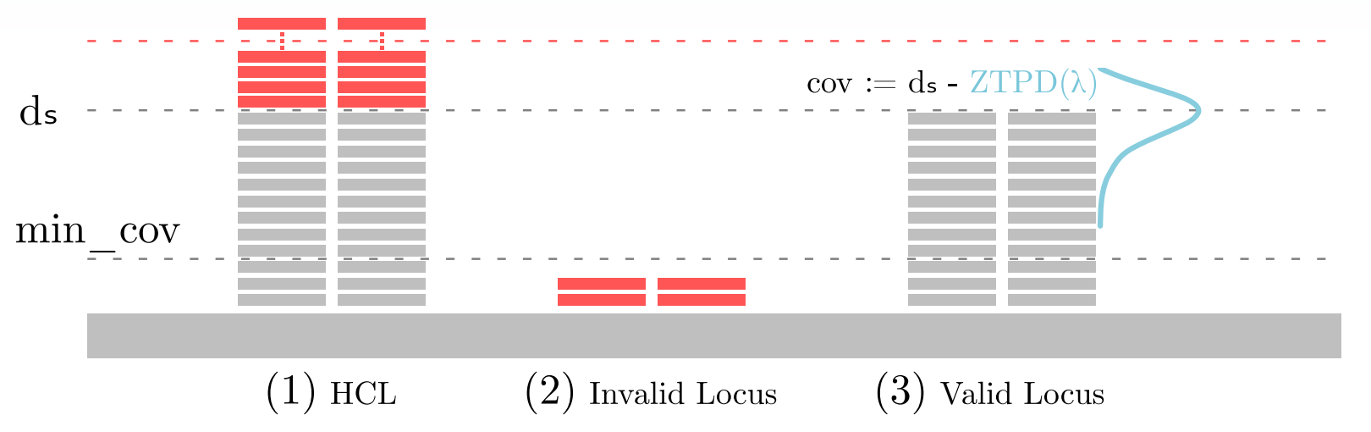
PCR duplicates, generated in the sequencing process, hold no information for analysis and have to be removed. In this simulation step, degenerate base regions (DBRs) [Schweyen et al. (2014)], pseudo-unique identifiers, have already been assigned to each read. Hence, PCR duplicates are added as carbon copies of reads, thereby creating duplicate DBRs.

Sequencing Errors (Step 8)

Using the Illumina error model, each base has a chance of $p = 0.01$ to be called incorrectly. If a base is changed, this is reflected in the simulated base qualities by a low quality value. The p7 barcode, though not part of the read itself, can be affected as well.

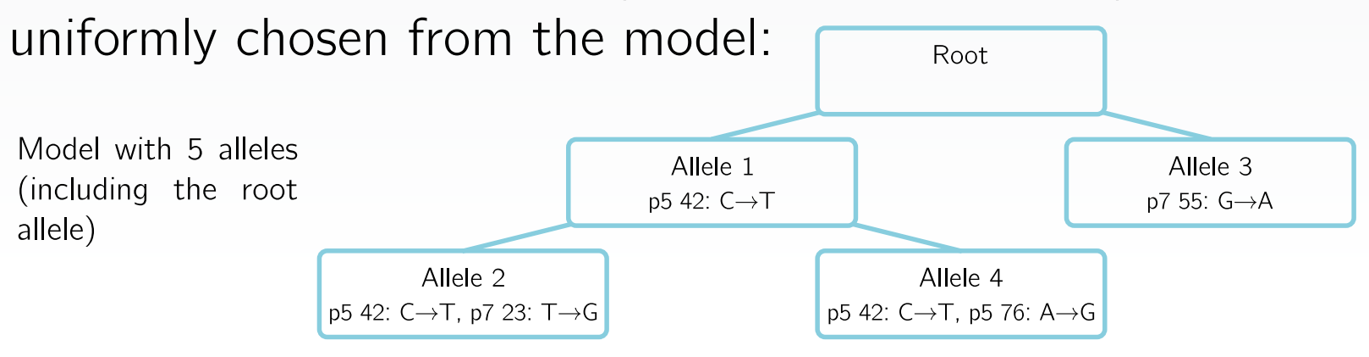
Coverage

Since the target sequencing depth d_s is seldomly reached, the coverage is modified using a PD to create a realistic coverage behavior. The given d_s is used as mean of the distribution:



SNP Model

SNPs are simulated using a tree based model in which each allele receives one SNP more than its parent allele and the root allele is the template sequence without SNPs. A SNP can occur on p5 or p7 side of the locus. The number of different alleles simulated is chosen from a ZTPD, enforcing at least one SNP per locus.
For each SNP event, one (or two for heterozyg.) allele is uniformly chosen from the model:



Poisson-Distribution (PD)

Discrete probability distribution, modeling number of events for a given interval.

Parameter: $\lambda :=$ expected #events per interval, $\lambda \in \mathbb{R}_{\geq 0}$

Mean: λ PMF: $P(X = k; \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$

Zero-Truncated PD (ZTPD)

Discrete probability distribution, modeling number of events for a given interval and >0 events.

Parameter: $\lambda :=$ expected #events per interval, $\lambda \in \mathbb{R}_{\geq 0}$

Mean: $\frac{\lambda e^{\lambda}}{e^{\lambda}-1}$ PMF: $P(X = k; \lambda) = \frac{\lambda^k}{(e^{\lambda}-1)k!}$

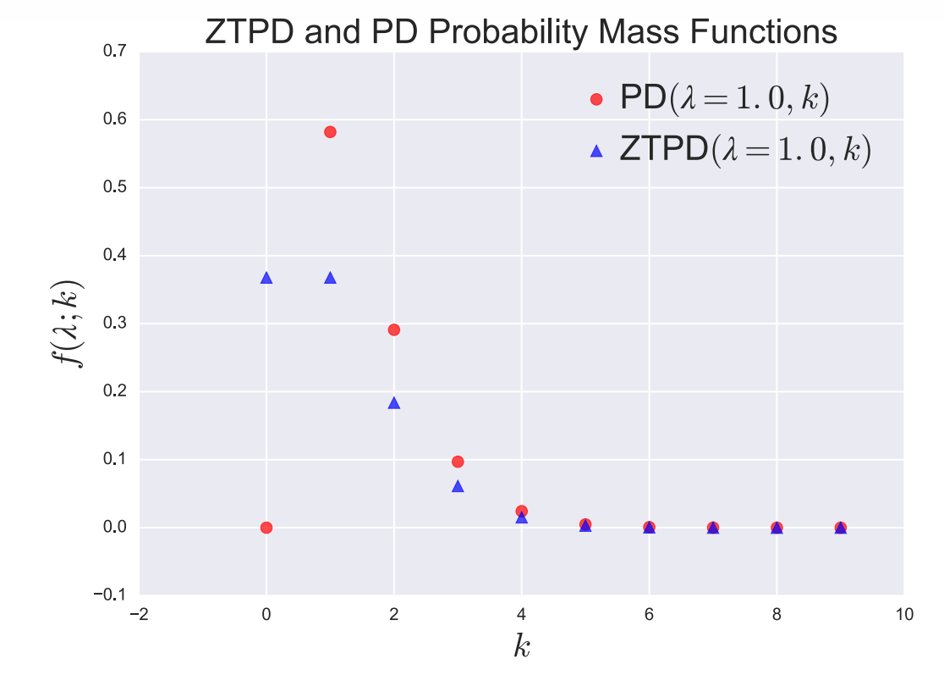


Figure: PMFs
PMFs of PD (red dots) and ZTPD (blue triangles) for λ = 1.0 and k = 0..9.

Intuitively:
Selecting values from a ZTPD behaves like using a PD but rerolling all x = 0.

Choosing from a (ZT)PD:
Evaluate the CDF of the target distribution for a random value 0 <= p < 1.

Ground Truth

RAGE creates logs and statistics describing the data set and a verifiable ground truth, containing all detectable simulated effects. Three different data formats are used:

Stacks TSV: Common and easy to convert format; single-end only: breaks up p5 and p7 info into two files, decoupling genotype information.

RAGE GT: Paired-end format providing more detail, including: coverage, allele frequency and p5+p7 genotypes.

Annotated FASTQ: All modifications (SNPs, PCR copies, sequencing errors, etc.) are also added to the FASTQ name lines.

Availability

RAGE is being developed in Python 3 and Cython. It will be available in the fourth quarter of 2016 under MIT License on github, PyPI and bioconda.

References:

Catchen, J. et al. (2013). Stacks: an analysis tool set for population genomics. Molecular Ecology, **22**(11), 3124–3140.
Schweyen, H. et al. (2014). Detection and Removal of PCR Duplicates in Population Genomic ddRAD Studies by Addition of a Degenerate Base Region (DBR) in Sequencing Adapters. The Biological Bulletin, **227**(2), 146–160.
Mora-Márquez, F. et al. (2016). ddRADseqTools: a software package for in silico simulation and testing of double digest RADseq experiments. Molecular Ecology Resources.
Lepais, O. et al. (2014). SimRAD: an R package for simulation-based prediction of the number of loci expected in RADseq and similar genotyping by sequencing approaches. Molecular Ecology Resources, **14**(6), 1314–1321.
Eaton, D. (2014). PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. Bioinformatics, **30**(13), 1844–1849.