**Introduction**

This readme describes the HMM used for identifying genomic introgressions described in “Fine scale mapping of genomic introgressions within the *Drosophila yakuba* clade”**,** Turissini and Matute (submitted)**.**  The programs are presented as is and are not being actively supported. Scripts will need to be edited to contain connection information (user, password, database, hostname) for MySQL. Also, several sections are hardcoded to use the 5 major chromosome arms found in *Drosophila* (*2L, 2R, 3L, 3R, X*). If other chromosome names are required, the programs will need to be edited accordingly.

**Prerequisites**

* Files
  + Filtered vcf file containing genotype calls from GATK for individuals from both the donor and recipient species
* Perl
  + Modules
    - DBI
    - DBD::mysql
* Python 2
  + Modules
    - \_\_future\_\_
    - datetime
    - random
    - math
    - sys
    - MySQLdb
    - os
    - warnings
* Cluster running LSF
* MySQL
  + All scripts will need to be edited to contain connection information
  + The following tables need to be created
    - introgression\_snps\_params

+--------------+--------------+------+-----+---------+-------+

| Field | Type | Null | Key | Default | Extra |

+--------------+--------------+------+-----+---------+-------+

| run\_name | varchar(30) | NO | PRI | NULL | |

| group\_list | varchar(255) | YES | | NULL | |

| arms\_table | varchar(30) | YES | | NULL | |

+--------------+--------------+------+-----+---------+-------+

* run\_name is a descriptive name for the hmm run
* group\_list is a string containing the names of the group\_names from tracking\_groups that will be part of the run. The list should be in the form 'A', 'B', 'C'
* arms\_table contains the name of the table $arms\_table described below
  + - tracking\_bams

+-----------------+---------------+------+-----+---------+----------------+

| Field | Type | Null | Key | Default | Extra |

+-----------------+---------------+------+-----+---------+----------------+

| tb\_id | int(11) | NO | PRI | NULL | auto\_increment |

| ind | varchar(100) | YES | MUL | NULL | |

| bam\_file | varchar(255) | YES | | NULL | |

| cov\_99\_quantile | int(11) | YES | | NULL | |

+-----------------+---------------+------+-----+---------+----------------+

* tb\_id is an auto\_increment integer
* ind is the name of the individual
* bam\_file is the path to the bam file containing the mapping results for ind but can be any unique name
* cov\_99\_quantile is the 99th quantile of the per site coverage distribution from the bam file
  + - tracking\_groups

+------------+--------------+------+-----+---------+----------------+

| Field | Type | Null | Key | Default | Extra |

+------------+--------------+------+-----+---------+----------------+

| tg\_id | int(11) | NO | PRI | NULL | auto\_increment |

| ind | varchar(100) | YES | MUL | NULL | |

| group\_name | varchar(50) | YES | MUL | NULL | |

+------------+--------------+------+-----+---------+----------------+

* tg\_id is an auto\_increment integer
* ind is the name of the individual
* group\_name is the name of the group the ind belongs to. The group names should refer to the donor and recipient species
  + - tracking\_ref\_genomes

+----------------------------+--------------+------+-----+---------+-------+

| Field | Type | Null | Key | Default | Extra |

+----------------------------+--------------+------+-----+---------+-------+

| ref\_genome | varchar(255) | NO | PRI | NULL | |

| ref\_db\_table | varchar(50) | YES | | NULL | |

| chrom\_name\_field | varchar(20) | YES | | NULL | |

| chrom\_len\_field | varchar(20) | YES | | NULL | |

+----------------------------+--------------+------+-----+---------+-------+

* ref\_genome can be the path to a file containing the reference genome or just a unique name
* ref\_db\_table should be the name of the table described below as $arms\_table
* chrom\_name\_field is the name of the chromosome name field in $arms\_table which is arm in the example below
* chrom\_len\_field is the name of the field containing the chromosome length in $arms\_table
  + - tracking\_vcf\_bams

+----------+--------------+------+-----+---------+----------------+

| Field | Type | Null | Key | Default | Extra |

+----------+--------------+------+-----+---------+----------------+

| tvb\_id | int(11) | NO | PRI | NULL | auto\_increment |

| vcf\_file | varchar(255) | YES | | NULL | |

| ind | varchar(30) | YES | MUL | NULL | |

| bam\_file | varchar(255) | YES | | NULL | |

+----------+--------------+------+-----+---------+----------------+

* tvb is an auto\_increment integer
* vcf\_file is the name of the vcf file passed as a parameter into get\_site\_all\_introgress\_pop\_from\_vcf.pl
* ind is the individual name
* bam\_file should be the same bam\_file associated with ind in tracking\_bams. A record should be present for every ind in the vcf file.
  + - $repeat\_regions\_table
      * Contains locations of repetitive regions of the genomes which are used for filtering spurious introgressions

+-------+------------+------+-----+---------+----------------+

| Field | Type | Null | Key | Default | Extra |

+-------+------------+------+-----+---------+----------------+

| rr\_id | int(11) | NO | PRI | NULL | auto\_increment |

| arm | varchar(5) | YES | | NULL | |

| start | int(11) | YES | MUL | NULL | |

| end | int(11) | YES | MUL | NULL | |

+-------+------------+------+-----+---------+----------------+

* rr\_id is an auto\_increment integer
* arm is the chromosome name
* start is the start of the repetitive region
* end is the end of the repetitive region
  + - $arms\_table
      * contains the chromsomes present in the genome

+---------+------------+------+-----+---------+-------+

| Field | Type | Null | Key | Default | Extra |

+---------+------------+------+-----+---------+-------+

| arm | varchar(6) | NO | PRI | NULL | |

| arm\_len | int(11) | YES | | NULL | |

+---------+------------+------+-----+---------+-------+

* arm is the chromosome name and any name can be used as long as the field name is also in tracking\_ref\_genomes
* arm\_len is the chromosome length and any name can be used as long as it is also in tracking\_ref\_genomes.

**Running the HMM pipeline**

1. **get\_site\_all\_introgress\_pop\_from\_vcf.pl** – This program identifies the markers to be used by the HMM. Markers are stored in a separate table for each arm. It takes the following input parameters:
   1. $prefix – the prefix for the run. Note that this should not exceed 7 characters (and maybe less if there are long ind names) such that database tables created by later programs do not exceed the MySQL length limit.
   2. $run\_name – the name of the run\_name in the introgression\_snps\_params table that contains the group\_list and arms\_table for this run.
   3. $donor\_group – The group from which the introgression is coming from.
   4. $recip\_group – The group into which the introgression is going.
   5. $freq\_cutoff – The allele frequency cutoff for markers such that abs(donor\_freq – recip\_freq) >= $freq\_cutoff
   6. $repeat\_regions\_table – database table containing the unique repetitive regions for the ref genome the reads were mapped to.
   7. $gatk\_vcf – vcf file containing genotypes for the groups present in the group\_list from introgression\_snps\_params.
2. **hmm\_introgress\_all\_downsample\_single\_arm\_master.pl** – The master program submits individual jobs to run the HMM for each arm/ind combination on a cluster using bsub. The HMM results for each ind are stored in their own database table. The program takes the following input parameters:
   1. $prefix – same prefix as 1a.
   2. $recip\_group – same recip\_group as 1d.
   3. $pab – sequencing error parameter for hmm: use .01.
   4. $base\_a\_prob – per base recombination rate used for transition probabilities: use 1e-9
   5. $err\_multiplier - multiplier for per base recombination rate used for transition probabilities involving error states: use 25000
   6. $freq\_diff\_cutoff – same as 1e.
3. **hmm\_genotype\_tracks\_master\_parallel.pl** – Tracks are contiguous blocks of SNPs with the same most probable genotype as determined by the HMM. This master program submits a job for each ind on a cluster using bsub. The tracks program makes unfiltered and filtered tracks tables as well as a filter log table that is only used for debugging purposes. The program takes the following input parameters:
   1. $prefix - same prefix as 1a.
   2. $suffix – suffix identifying how markers were chosen: use all\_30
   3. $hmm\_type - a 3 character name identifying the hmm program that was run. Will be ‘dsm’ most of the time.
   4. $recip\_group – same as 1d.
   5. $repeats\_table - same table as 1f.
4. **make\_hmm\_introgress\_regions.pl** – The regions program should be run after all of the tracks jobs have completed. It creates a series of summary tables that contain information for all of the inds. The unftrk table contains info for all unfiltered tracks. The tracks table contains all of the filtered tracks. The regions table contains all of the regions defined as a set of overlapping tracks. The regions\_inds table contains all of the inds with an introgressed track for each region. The blocks table contains blocks which are regions of the genome with the same number of inds with introgressed tracks. The blocks\_inds table contains all of the inds with introgressed tracks for each block. The program takes the following input parameters:
   1. $prefix - same prefix as 1a.
   2. $introgress\_group – same as 1d.
   3. $hmm\_suffix – combines info from pab, err\_multiplier, and suffix: use 01\_9\_25000\_all\_30 if using the suggested parameters
   4. $hmm\_type – same as 1c.
   5. $repeats\_table - same table as 1f.

**Interpreting the results**

Results can be found in the consolidated tracks table created by make\_hmm\_introgress\_regions.pl with the name [prefix]\_hmm\_dsm\_pab\_[hmm\_suffix]\_tracks. The genotype field will contain 3 values:

* homo\_1 – homozygous for the recipient species allele
* het – heterozygous for the donor and recipient species alleles indicative of a heterozygous introgression
* homo \_2 – homozygous for the donor species allele indicative of a homozygous introgression

introgress\_snps is the count of snps within the track that appear to be introgressed. per\_repeat is the percentage of the track that overlaps with repetitive sequence.