Microsatellite development using Galaxy:
User manual (v1.2)

Table of contents

1) Introduction

This document is to support users of the Galaxy-based microsatellite development pipeline described in

> Griffiths et al. (2016) A Galaxy-based pipeline for optimized, streamlined microsatellite development from Illumina next-generation sequencing data. Conservation Genetics Resources.

Please read this before using any of the tools described. We also advise that you consult the original research papers for pre-existing programs used in the pipeline (see pg. 24).

This manual describes in detail how to use Palfinder Galaxy Service (https://palfinder.ls.manchester.ac.uk) to develop microsatellite markers, implementing the methods of Griffiths et al. (2016). This will be the most appropriate option for the majority of users and is freely available. However, we have also included instructions for using the programs at the command line (Ubuntu), and for those with a local Galaxy server at their institution.

2) Illumina paired-end sequencing

I- Platform and read length –

For our projects, we have used the Illumina MiSeq to create 2 x 250 base pair (bp) reads or 2 x 300 bp reads. However, Castoe et al (2012) also found the GAIIx an effective platform, with read lengths of 2 x 116bp. There is likely to be a trade off between read length and cost - longer reads are generally more expensive but the longer the read, the more chance of finding microsatellite loci with suitable flanking region for primer design. It is also important to note that read quality significantly reduces at the end of reads in most sequencing projects, and therefore after quality filtering and trimming, the reads used for microsatellite development will likely be shorter than the original sequencing length.

II-Sample multiplexing –

Multiple samples can be run in a single flowcell of an Illumina sequencer, which reduces the cost-per-species of sequencing, and therefore the microsatellite development. However, this is also a trade off between cost/number of species and depth of coverage/number of reads achieved for each species. In our lab we have multiplexed up to eight samples (from eight species) on a single flowcell – please see Table 1 in Griffiths et al (2016) for the numbers of microsatellites and primers found with various filtering combinations.

As sequencing technology is constantly evolving, we would recommend consulting a sequencing technician at your chosen facility for their advice on both platform and multiplexing potions.

3) Microsatellite development using Galaxy

I) Palfinder Galaxy Service: getting started

a) Introduction

Palfinder Galaxy Service is a preconfigured version of Galaxy (an open source bioinformatics tool for handing next-generation sequencing data) with all the programs and abilities detailed in the bioinformatics pipeline described in Griffiths et al. (2016). Palfinder Galaxy Service can be accessed at https://palfinder.ls.manchester.ac.uk. This service provided by the Bioinformatics Core Facility and Research Computing teams at the University of Manchester, in conjunction with the lab of Richard Preziosi (http://www.preziosilab.org). Palfinder Galaxy Service allows users to:

- Create a user account
- Upload Illumina sequence data files
- Use example data files
- Run a microsatellite development pipeline using the following tools within the Galaxy environment:
  - FastQC
  - Trimmomatic
  - Pal_finder (also utilises Primer3, Pal_filter and PANDASeq)
- Download results files
- Store data for 1 day.

b) Creating an account

On the Galaxy Palfinder Service homepage (https://palfinder.ls.manchester.ac.uk) please create a free user account. You will be required to verify your account through email. Please read the terms and conditions of use of the service before creating an account. By registering an account you agree to be bound by these terms and conditions.

c) Basic information and navigation

Using Galaxy is simple and intuitive, and the main Galaxy website https://galaxyproject.org has many tutorials for first-time users. We recommend you take a look at these in addition to these basic starting points:

- When you enter Galaxy, there are three main sections to the screen. The left hand column contains a list of the tools that you can use; the middle column is where options can be configured and data/results viewed; the right hand column is your 'history' – this shows the output files for all the jobs you have run (green), or the jobs that are currently running (yellow) or any failed jobs (red). You can create new histories and switch between histories using the cog icon on the right hand side. Available memory for your user account is shown in the top right hand corner (quota is 20GB).

- Running jobs – When you run a job on the Galaxy Palfinder Service, it does not use your computer's own processing power to run the job – the website acts as an interface while the job is actually run using servers at the University of Manchester. This means that while your job is running you can use your computer as normal, close your browser window, shut down your computer etc. You can then log back into your account and check the progress of the job when you are ready – when the output files have turned green in your history, the programs have finished running and the files are ready to be viewed. You may also queue jobs – i.e., if you are currently running a job that will generate a dataset with which you want to run further programs on, you can set up the next job which will run automatically when the required file is ready.

- Deleting data from your history – When you delete a dataset from your history using the 'x' icon on the right, the command is analogous to deleting files on your computer when they are sent to the recycling bin instead of being permanently deleted. As such, your memory usage will not immediately decrease after deleting an item from your history. To delete permanently, click 'purge deleted datasets'. After this is carried out, the file is irretrievable so use caution.

- **All data (input and output files) will be deleted permanently after 1 day, so please download and save all files you need to your personal computer!**

d) Uploading data

Select 'Upload data from your computer' in the Tools list on the left hand side of the screen. Users can upload fastq data files up to 2GB directly to the server via a web browser. Larger files can be

uploaded via FTP server at ftp://palfinder.ls.manchester.ac.uk using your email address and Galaxy password to log in. (Uploaded files will then be available for import via the Upload tool.)

e) Example data files

Input files:

We have provided example data files – please feel free to experiment with the programs using these. Example input files consist of two files (R1 and R2, as it is paired end data) containing 250bp-length reads generated by the Illumina MiSeq. Note that this is a very small subset of reads – real data files are likely to take much longer to run, and yield a much higher number of microsatellites and primers!

Example data files can be accessed through the Shared Data link at the top of the page, in 'Data Libraries'.

Example history:

We have made an example 'history' showing the processes and outputs of the full pipeline. This can be accessed through the following pathway:
'Shared data' → 'Published histories' → 'Palfinder example'.

See: 4) Running the pipeline in Galaxy (including Palfinder Galaxy Service) (pg. 9) for instructions in running the pipeline

II) Installing pipeline programs onto local Galaxy servers

Note: Ignore this section if you are using the Palfinder Galaxy Service!

If your institution has a local Galaxy server, you may download the wrappers for FastQC, Trimmomatic and Pal_finder (including Primer3, pal_filter and PANDASeq). via the Galaxy Tool Shed (https://toolshed.g2.bx.psu.edu) onto your local Galaxy server:

FastQC - https://toolshed.g2.bx.psu.edu/view/devteam/fastqc

Trimmomatic - https://toolshed.g2.bx.psu.edu/view/pjbriggs/trimmomatic

Pal_finder (including Primer3 and Pal_filter and PANDASeq for optimal loci filtering and paired read assembly) - https://toolshed.g2.bx.psu.edu/view/pjbriggs/pal_finder.

Instructions for installation can be accessed at https://github.com/fls-bioinformatics-core/galaxy-tools/tree/master/tools/pal_finder/.

See: 4) Running the pipeline in Galaxy (including Palfinder Galaxy Service) (pg. 9) for instructions in running the pipeline

4) Running the pipeline in Galaxy (including Palfinder Galaxy Service)

Using Galaxy (whether your own local server or the Palfinder Galaxy Service) the pipeline can be run in 4 steps, outlined below. We stress the need to consult the original literature cited for more details about these programmes, their configuration options, and the results they output. Note that we have simplified some of these programs in Galaxy by removing some of the options we considered irrelevant for microsatellite development purposes.

I - Quality assessment of raw data (FastQC)

It is useful to check the quality and characteristics of the raw sequence files. In the left-hand pane of the Galaxy window select the FastQC option. Select one of your data files and click 'execute'. You will see the job running in the 'History' pane on the left hand side of the window (it will be yellow while it is running, and green when the job is complete). While the job is running, repeat for the other file. This step should be fairly rapid, and the output will give you basic quality information and stats on your raw data. Each FastQC run will generate two output files – the 'Webpage view' and the 'Raw data' file. The 'webpage view' shows the information in useful graphics, but the raw data file can be viewed for the full break down of figures.

II - Filtering and trimming of reads (Trimmomatic)

In the list of options in the left hand pane, select Trimmomatic. You should select your data files from the drop-down menu for each side of the read (taking care to select the file that corresponds to the correct option [i.e., R1 and R2]). There are a number of options to configure within Trimmomatic (see Bolger et al [2014] and the Trimmomatic user manual <http://www.usadellab.org/cms/index.php?page=trimmomatic>). The 'default' settings that we use result in files containing reads with Phred scores of 20 (i.e., ≥99% base call accuracy) and minimum lengths of 50bp. These settings are: SLIDING WINDOW= 4bp window size, quality score =20, LEADING= 4, TRAILING= 4, and MINLEN =50. Simply select 'add function' and then pick your chosen function from the drop-down menu. Note that the order you put the functions matter as Trimmomatic will go through this sequentially! Adapter contamination may also be removed using the ILLUMINACLIP step (check the output from FastQC to check the presence and identity of any adapter sequences in your data). When you have selected the functions you wish to use, click 'execute'. Trimmomatic will generate four output files – two files containing the surviving pairs [R1 paired] and [R2 paired]), and two files containing reads in which one or both reads in a pair did not both pass the quality threshold [R1 (unpaired)] and [R2 (unpaired)].

III - Quality assessment of trimmed data (FastQC)

To check the quality and basic stats of your trimmed data sets, repeat the FastQC step as above using the [R1 (paired)] and [R2 (paired)] files generated by Trimmomatic. If this is satisfactory, proceed to the next step. If not, adjust the Trimmomatic settings and re-run.

IV - Microsatellite isolation, primer design, read assembly and loci filtering (pal_finder)

Click 'pal_finder' in the tools list. Input the [R1 (paired)] and [R2 (paired)] files generated by Trimmomatic in the drop down lists.

a – Filtering options

There are 3 tick-box options which will filter the output from Pal_finder and put it in a new tab delimited file called 'Filtered Microsatellites – full details' (whilst retaining the original file). This function makes it easier to view only the loci that the user is interested in, as there are often thousands of results in the Pal_finder output.

- *Only include loci with designed primers*

Many loci are found which have insufficient or unsuitable flanking regions for primer design. This option if selected will only show those loci for which Pal_finder could design primers for in the 'Filtered Microsatellites – full details' output file.

- *Exclude loci where the primer sequences occur more than once in the reads*

Pal_finder scans the entire set of reads for the primer sequences of each primer pair. If they occur more than once, this may be because there the copy number is more than one, i.e., the primer sequence does occur elsewhere in the genome of the organism. If so, the primer may bind to non-target regions in PCR resulting in non-specific amplification. However, another cause could be that the same region of DNA has been sequenced more than once and there are multiple reads containing the same region. By selecting this option, the filtered output file will only include loci in which the forward and reverse primer sequences have only occurred once in the entire set of reads.

- *Only include loci with 'perfect' motifs, and rank by motif size:*

Users can chose to remove imperfect, interrupted and compound repeats from the filtered output (e.g., $AC_{(16)}N_{(8)}TG_{(9)}$; $ATCT_{(17)}GT_{(8);}$ $GC_{(14)}GC_{(16)}$). These types of microsatellites do not follow the stepwise mutation model and therefore some researchers may wish to avoid

them. When this is selected, only perfect repeating units (e.g., $TC_{(12)}$; $ACTC_{(24)}$; $AGG_{(11)}$) are shown in the filtered output file.

b - Read assembly

Below the tick-box filtering options, there is a Yes/No option for:

'Use PANDAseq to assemble paired-end reads and confirm primer sequences are present in high-quality assembly'.

If selected, each pair of reads are assembled individually using PANDAseq (Masella et al. 2012) software (PEAR assembly algorithm [Zhang *et al.* 2014, minimum overlap = 25nt, confidence score = 95] to produce a single high-quality assembly. The two primer positions are then identified within this assembly, confirming that both primer sequences occur in the same region of DNA template. When this option is selected, only loci resulting from reads that have been successfully assembled will be shown in the 'Assembly' output file, along with any filters applied in the previous step.

c – Microsatellite type and repeat unit specification

In the next set of configuration options, users can specify the minimum number of repeats for Pal_finder to look for in each microsatellite type (i.e., 2-mer, 3-mer, 4-mer, 5-mer or 6-mer). Setting any of these as '0' means that Pal_finder will not search for this type of repeat unit. The more repeated units, the more variable the locus is likely to be (our default option is to use a minimum of 8 repeats).

d – Mispriming library to use

This option allows specification of sequences or interspersed repeats that Primer3 should avoid as an area for primer design. We recommend using the default library for any non-model organisms, however you may also use a library of custom sequences.

e - Primer design

There is a drop-down menu to select either the default settings for Primer3 for your primer design, or to customise them. We would recommend customising the settings to comply with the specifications of the PCR reagents and conditions you plan to use for primer testing and future genotyping – for example, we use the Qiagen Type-it® Microsatellite PCR Kit, which specifies a number of recommended parameters for primers for optimal performance of the kit. If you want to use multiplex PCR to increase efficiency and decrease costs of genotyping, primers should be designed

with similar melting temperatures to enable them to successfully amplify using the same cycling conditions.

f - Config file

There is a Yes/No option to ‘Output the config file to the history’. This is the configuration file for the Pal_finder part of the process, and shows the input parameters and the script that can be run at the command line.

After all the settings have been configured, click ‘execute’. The amount of time that this will take to run depends on the size of your data files and the options chosen, but may be over 24 hours.

5) – Output

All or a selection of the following files will be outputted into your history, depending on the options selected in the previous step:

I - All microsatellites – full details: This file shows all the loci and primers that Pal_finder has located using the parameters specified. This file is unfiltered, so the output is the same regardless of if any of the additional filter options were selected – all loci and primers are shown, with detailed information including read IDs, primer sequences and occurrences of the primer sequences in the reads. We recommend downloading and importing this tab delimited file into Microsoft Excel. By using the column filtering function in Excel, the results can be managed more easily to show desired microsatellites.

II - Filtered microsatellites – full details. This file gives all the same information as in the previous file, but only shows the loci that have passed any of the three filters applied by users. Note that the loci in this file have not passed the 'assembly' stage too if this option was selected (please see Assembly output file for these loci). Again, this is a tab delimited file that can be imported into Microsoft Excel for ease of use.

III- Assembly – If you selected the assembly option, loci found in assembled reads are shown in this file. Any filters that have been selected also apply to this file. So for example, if all three filters and the assembly have been selected, the loci shown in this file will: 1) result from successfully assembled reads, 2) be perfect repeats, ranked by size, 3) will all have primers designed for them, and 4) have primer sequences that only occur once in the entire set of reads. Loci in this file are the result of the most stringent selection process, and in our lab we will test the loci found in this file first. If there are not enough loci in this file, we will then chose loci from the 'Filtered Microsatellites – full details' file that did not occur in the 'Assembly' file as well. If this does not yield enough loci, we then consult the 'All Microsatellites – full details' file.

IV - Summary of microsatellite types - This file gives you basic information about the number and type of microsatellite loci found by Pal_finder.

V - Config file (pal_finder) – this can be adequately viewed within the middle Galaxy results window or can be downloaded as a .txt file.

6) Running the pipeline from the command line

Applicable to Ubuntu (and variants) only.

Indented text should be entered in the terminal on your system.
Example command:

    this/in/an/example/command

In this document a command may fall across several lines but should be entered as a single command.

You may be able to run the pipeline on Windows under the Cygwin environment (https://www.cygwin.com/) but this is entirely untested.

You will need admin rights / sudo access for this process.

Warning! These instructions are for advanced users only. Please read through the entire instruction set before beginning. If in doubt, do not modify your system based on these instructions. We accept no liability for damage to, or loss of data from your system as a result of following these instructions. We also give no guarantee that these instructions will work on your system.

Install and run FastQC

Download and install the FastQC package from:
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Install OpenJDK and FastQC using:

    sudo apt-get install openjdk-8-jre && sudo apt-get install fastqc

Run FastQC on your sequence files using the following command (substituting your own path and filename for "*/path/to/demo_file.fastq*"

    fastqc /path/to/demo_file.fastq

This will produce a FastQC report in html format in the same folder as your sequence data. View this file in an internet browser such as Firefox (www.mozilla.org)

Install and run Trimmomatic

Download Trimmomatic from http://www.usadellab.org/cms/?page=trimmomatic

Select and download the 'binary' package, unzip the file and note the location of the trimmomatic-0.36.jar (a newer version may exist since writing).

Run Trimmomatic using the following command (substituting your own paths and filenames for where appropriate.

```
java -jar path/to/trimmomatic-0.36.jar PE path/to/input_R1.fastq path/to/input_R2.fastq
output_forward_paired.fastq output_forward_unpaired.fastq
output_reverse_paired.fastq output_reverse_unpaired.fastq LEADING:3 TRAILING:3
SLIDINGWINDOW:4:20 MINLEN:50
```

You will also need to define the -phred and -ILLUMINACLIP arguments depending on your particular sequence data. Speak to your sequencing centre.

Install and run pal_finder

A prerequisite of pal_finder is that you have primer3 installed. Install it with the following command:

```
sudo apt-get install primer3
```

Download pal_finder_v0.02.04.tar.gz from

https://sourceforge.net/projects/palfinder/

Decompress the .tar.gz file using:

```
tar -xzf pal_finder_v0.02.04.tar.gz
```

Move into the pal_finder_v0.02.04 folder and note the location of the pal_finder_v0.02.04.pl and config.txt files.

Make a copy of “config.txt” naming it “config_working_file.txt”. Open “config_working_file.txt” using a plain text editor (such as gedit, nano or atom). An 'office' word processor such as Microsoft Office or LibreOffice Writer should not be used.

In the config_working_file, edit the following parameters (replace the italicised text below)

inputReadFile */path/to/input_R1.fastq*

pairedReadFile */path/to/input_R2.fastq*

MicrosatSumOut */path/to/output/microsat_summary_file.txt*

PALsummaryOut /path/to/output/PAL_summary_file.txt

2merMinReps *6*
3merMinReps *6*
4merMinReps *6*
5merMinReps *6*
6merMinReps *6*

primer3executable */path/to/primer3/executable*

PRIMER_OPT_SIZE *20*
PRIMER_MIN_SIZE *18*
PRIMER_MAX_SIZE *25*
PRIMER_MIN_GC *45*
PRIMER_MAX_GC *65*
PRIMER_GC_CLAMP *3*
PRIMER_MIN_TM *62*
PRIMER_MAX_TM *85*
PRIMER_OPT_TM *68*

Save the config_working_file.txt file.

You can then run pal_finder using the following command:

perl pal_finder.pl /path/to/config_working_file.txt

The pal_summary output from pal_finder is required as an input file for the pal_filter script below.

Install and run the pal_filter script

pal_filter requires that you have Biopython, python-dev and Pandaseq installed and can only be run under Python 2.7.

To install python-dev use the following command:

sudo apt-get install python-dev

To install Biopython, download the zip file from

https://github.com/biopython/biopython

Decompress the file using the command:

unzip biopython-master.zip

Move into the biopython-master folder and use the following command to install biopython:

python setup.py install

To install Pandaseq first install some required packages using the following command:

sudo apt-get install build-essential libtool automake zlib1g- dev libbz2- dev pkg-config

Download the zip file from

https://github.com/neufeld/pandaseq

Decompress the file using the command:

unzip pandaseq-master.zip

Move into the Pandaseq-master folder and install pandaseq using:

./autogen.sh && ./configure && make && sudo make install

Download the zip file from
https://github.com/graemefox/pal_filter

Decompress the file using:

unzip pal_filter-master.zip

Move into the pal_filter-master directory and make the script executable using the following command.

sudo chmod +x pal_filter.py

Run using the following command (substituting your own file paths where appropriate). This will turn on all the suggested filters to give (most stringent) results.

./pal_filter.py -i /path/to/input_R1_file.fastq -j /path/to/input_R2_file.fastq -p /path/to/pal_finder/pal_summary_file -primers -occurrences - rankmotifs -assembly

The -primers, -occurrences -rankmotifs and -assembly filters can all be disabled by simply leaving them out of the previous command.

The output from pal_filter.py is the final output of the workflow and contains details of microsatellite loci, primers and the raw reads from which they were optimised.

7) Next steps

The following is a brief guide to the process of microsatellite development beyond identifying suitable primers. The advice here is given in light of the experiences we have had in our lab gained in the course of many microsatellite development projects – use it at your own risk! This is an informal guide intended to assist those new to microsatellite development, and is by no means extensive - users should consult the literature also seek the help of those experienced with microsatellites if possible. However we hope this will be useful as a starting point.

1) Select loci and primers of interest. This may be dependent on the type of study that you wish to do. Some of the primers tested will most likely not amplify correctly, and therefore it is advised for researchers to test more loci than the final number you wish to have for your study. We have found success rates for microsatellite development to range from 30% to 70%, species (and luck!) dependent. Around 50% success rate seems to be the average when all filtering options are utilised and the assembly option is selected.

2) Order 'unlabelled' (i.e., without fluorescent tags) forward and reverse primers. We recommend ordering the lowest concentration available, as this is sufficient for many PCR reactions and will save money.

3) Test primers (unlabelled) for successful amplification in PCR. We test primers on 8 samples of DNA in singleplex PCR reactions (only one primer pair), and check the PCR products for successful amplification using agarose gel electrophoresis. To ensure adequate conservation of the primer binding regions across genetically diverse individuals, we recommend selecting samples from a diverse geographical range for this initial testing phase. Successful amplification can be classed as a single band or two close-together bands in a large majority of the wells. For the PCR, we use the Qiagen Type-it® Microsatellite PCR kit using the default cycling settings, however, we scale the reaction down to 5ul to reduce costs.

4) PCR optimisation (if necessary). Depending on the characteristics of your loci, primers and DNA, cycling parameters such as annealing temperature and cycle number can be modified, as well as concentrations of components in the PCR mix.

5) Test primers (labelled). The loci that have successfully amplified using unlabelled primer pairs can then go through to the next stage of testing. This involves fluorescently tagging the PCR product so that it can be sized using a capillary electrophoresis sequencer. This can be achieved by fluorescently tagging the forward primer at the 5" end with labels such as 6FAM, VIC, PET and NED (the labels chosen depend on the sequencer and the filter set that is used, so consult with sequencing technicians at your chosen facility before ordering primers. We have found 6FAM to be the most reliable florophore, and so we always use this for singleplex reactions.

   Alternatively, a 3-primer system may be used, which can save costs when genotyping at multiple loci. This involves tagging the forward primer with a 'tail' of a universal sequence at its 5" end, then adding the fluorescently tagged tail sequence alone as a third primer in the reaction in addition to the normal usual reverse primer. Purchasing multiple fluorescently-labelled forward primers can be expensive – this method reduces the number of fluorescently labelled primers that need to be purchased as the same tail sequence can be used for multiple primer pairs. We have had success with this method in our lab and recommend the papers Culley et al (2013) and Blacket et al (2012) for more information on this method.

   The fluorescently-labelled PCR products should be checked on a gel to ensure they have been correctly amplified (i.e., the bands are in the same position as when tested in the unlabelled reactions) before sending for genotyping using a capillary electrophoresis sequencer. As this fragment length analysis can also add significant cost to a microsatellite development project, a small number of samples per locus (we have used 8-10 in our projects) can be tested on initially to test their suitability. Software such as Genemapper or Peakscanner (ABI) can be used to view plots and to score alleles.

6) PCR optimisation (if necessary) and locus selection. It may be apparent from viewing traces that PCRs could be further optimised (e.g, there are additional peaks that suggest non-specific binding). A few loci may also turn out to be unsuitable (e.g., too many stutter bands) and may be excluded from further use.

7) Multiplex planning. By combining primers in multiplexes, costs and time spent on PCR and fragment length analysis can be reduced. By using the initial information on allele size ranges from the testing of the primers, multiplexes can be designed by labelling primers that produce

fragments of overlapping lengths with different florescent colours. Multiplex Manager (Holleley and Geerts 2009) can make this process very quick, easy and effective.

8) Multiplex testing. It is important to test your multiplexes are working correctly, firstly on a gel (you should see one or more bands, depending on the number of loci multiplexed and expected sizes), and then by fragment length analysis using a capillary sequencer. Some of the samples used to test multiplexes should be the same as in the singleplex testing to make sure that amplification is consistent (NOTE – if you are using the 3-primer system, changing the universal tail of a primer for an alternative tail of a different number of base pairs will increase or decrease the product length by the same number, so adjust accordingly).

9) Genotyping of ~30 individuals from a single location using all functional multiplexes. This will involve formally scoring and binning alleles (R package MsatAllele [Alberto 2009] is freely available for binning). Each locus must then undergo basic characterisation and screening. This includes calculating: 1) Observed heterozygosity; 2) Expected heterozygosity; 3) If there is significant deviation from Hardy-Weinberg Equilibrium; 4) If there are significant levels of linkage disequilibrium between loci; 5) Presence of null alleles. Loci may then be excluded from further testing – for example if there are two loci that appear to be linked, one must be excluded.

8) References

Alberto F (2009) MsatAllele_1.0: An R package to visualize the binning of microsatellite alleles. J Hered 100:394–7. doi: 10.1093/jhered/esn110

Blacket MJ, Robin C, Good RT, et al (2012) Universal primers for fluorescent labelling of PCR fragments--an efficient and cost-effective approach to genotyping by fluorescence. Mol Ecol Resour 12:456–63. doi: 10.1111/j.1755-0998.2011.03104.x

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics btu170–. doi: 10.1093/bioinformatics/btu170

Culley TM, Stamper TI, Stokes RL, et al (2013) An Efficient Technique for Primer Development and Application that Integrates Fluorescent Labeling and Multiplex PCR. Appl Plant Sci 1:1300027. doi: 10.3732/apps.1300027

Holleley CE, Geerts PG (2009) Multiplex Manager 1.0: a cross-platform computer program that plans and optimizes multiplex PCR. Biotechniques 46:511–7. doi: 10.2144/000113156

9) Citing the service, pipeline and programs

This manual has been written by Sarah M. Griffiths and Graeme Fox.

If you use the **Palfinder Galaxy Service** or the **microsatellite development pipeline described** please cite the following:

> Griffiths SM, Fox G, Briggs PJ, Donaldson IJ, Hood S, Richardson P, Leaver GW, Truelove NK & Preziosi RF (2016) A Galaxy-based bioinformatics pipeline for optimised, streamlined microsatellite development from Illumina next-generation sequencing data. Conservation Genetics Resources.

In addition, the programs used within should be cited as follows:

Pal_finder:

> Castoe T, Poole A, de Koning A (2012) Rapid microsatellite identification from Illumina paired-end genomic sequencing in two birds and a snake. PLoS One 7:e3095. doi: 10.1371/journal.pone.0030953.

Primer3:

> Koressaar T, Remm M (2007) Enhancements and modifications of primer design program Primer3. Bioinformatics 23:1289–91. doi: 10.1093/bioinformatics/btm091
>
> Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG (2012) Primer3 - new capabilities and interfaces. Nucleic Acids Res 40:e115. doi: 10.1093/nar/gks596

PANDASeq:

> Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD (2012) PANDAseq: paired-end assembler for Illumina sequences. BMC Bioinformatics 13:31. doi: 10.1186/1471-2105-13-31

Trimmomatic:

> Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-20. doi: 10.1093/bioinformatics/btu170

FastQC:

> Andrews S. (2010). FastQC a Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Instructions for citing **Galaxy** can be found here on the Galaxy website.