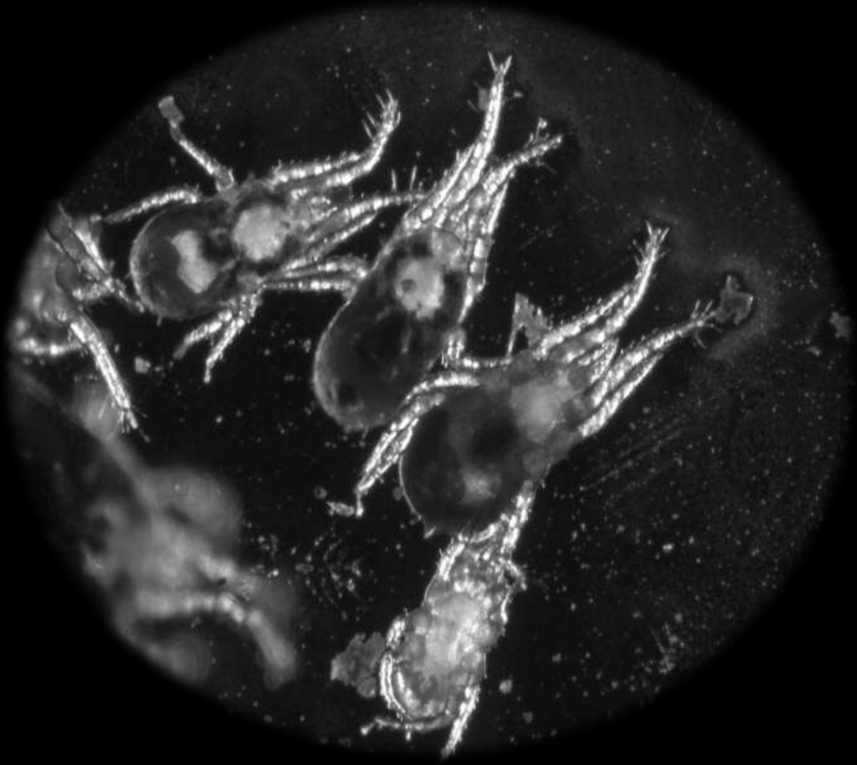
Utilising the genome analysis toolkit (GATK) to identify single nucleotide polymorphisms for use as genetic markers

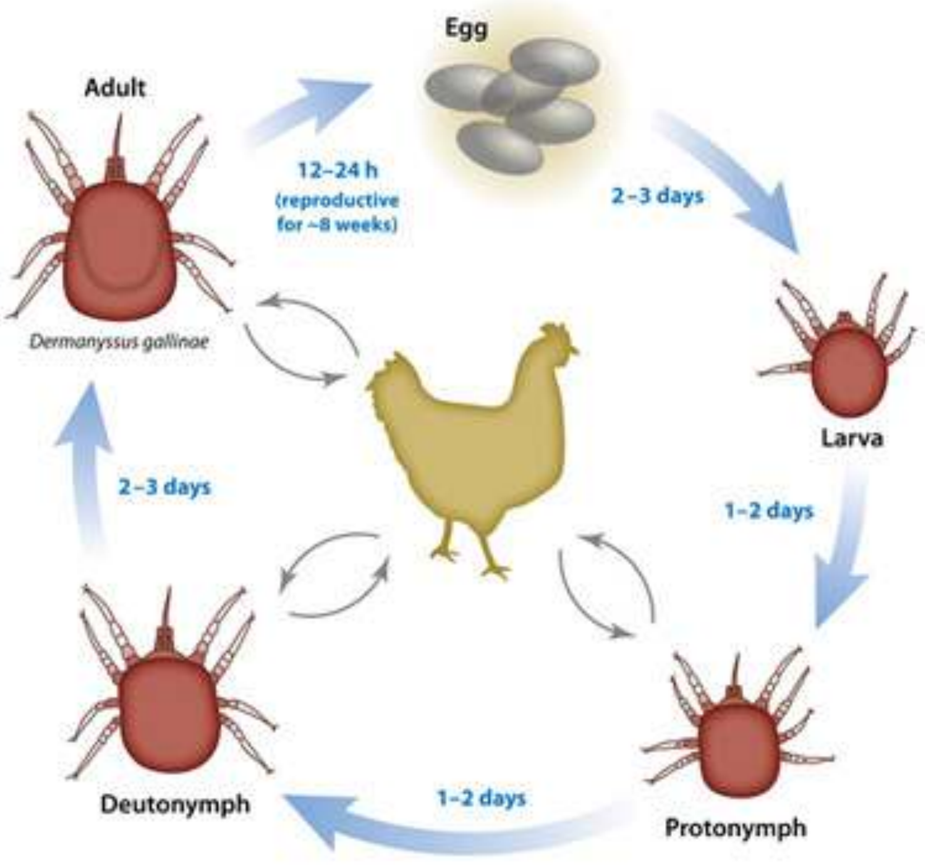
Eleanor Karp-Tatham

Supervisors: Prof Damer Blake, Prof Fiona Tomley, Dr Dong Xia & Dr Alasdair Nisbet (Moredun Institute)



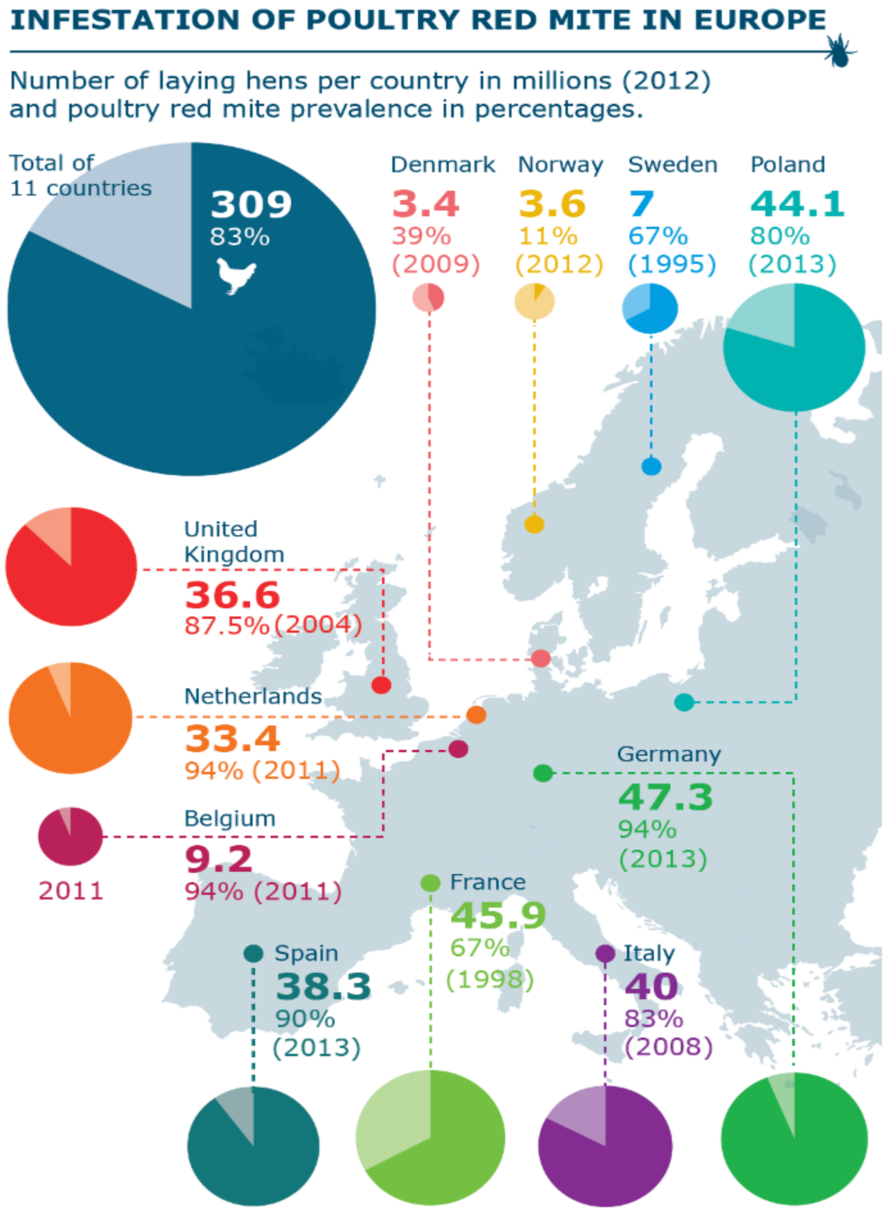
Dermanyssus gallinae

- Blood-feeding ectoparasite
- Significant welfare impact at moderate infestation levels
- €130 million loss to European poultry industry a year
- 28 avian hosts
- Vector for multiple pathogenic agents
- Five stage life-cycle
- Feed during hours of darkness (~30-90 minutes)
- Under optimal conditions (high relative humidity and 20-25°C) the complete cycle can occur in 7-10 days



Sparagano OAE, et al. 2014.
Annu. Rev. Entomol. 59:447–66

Current control of *Dermanyssus gallinae*

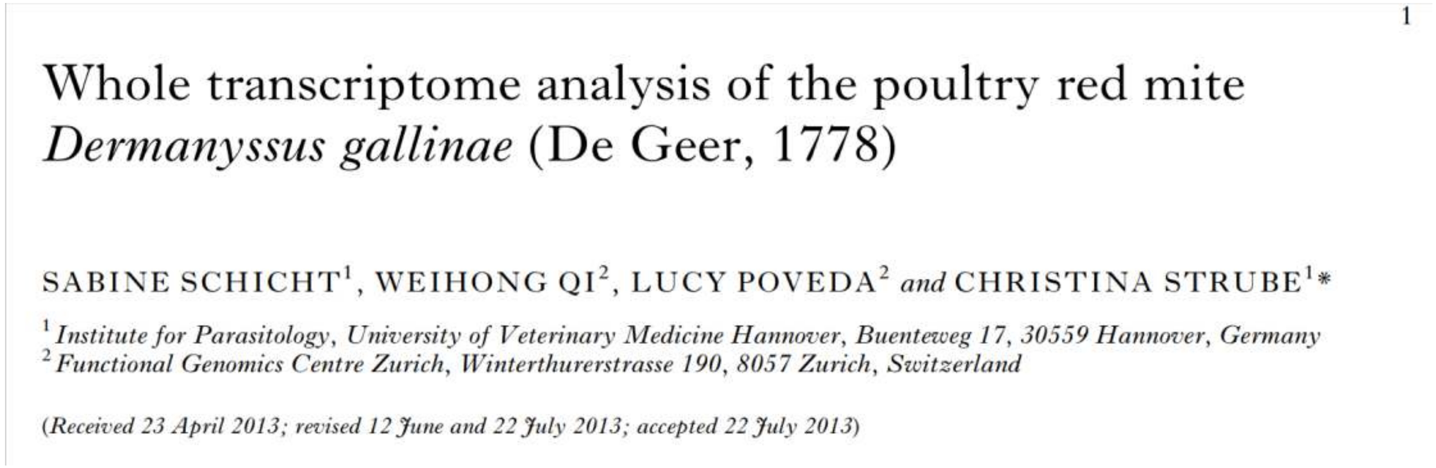


Genetics of *Dermanyssus gallinae*

ROYAL VETERINARY COLLEGE

- ➢ Transcriptomic data
- ➢ Illumina
- ➢ Not currently published

STIFTUNG TIERÄRZTLICHE HOCHSCHULE HANNOVER 1778



1

Whole transcriptome analysis of the poultry red mite *Dermanyssus gallinae* (De Geer, 1778)

SABINE SCHICHT[1], WEIHONG QI[2], LUCY POVEDA[2] *and* CHRISTINA STRUBE[1]*

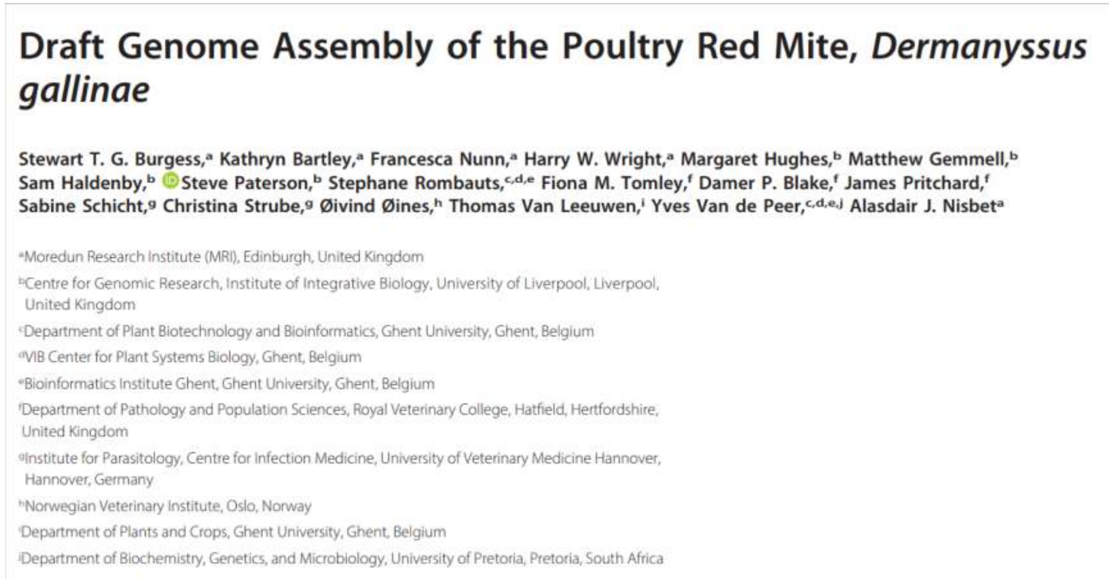
[1] *Institute for Parasitology, University of Veterinary Medicine Hannover, Buenteweg 17, 30559 Hannover, Germany*
[2] *Functional Genomics Centre Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland*

(*Received 23 April 2013; revised 12 June and 22 July 2013; accepted 22 July 2013*)

doi: 10.1111/mve.12050

ersity, 4-101 Koyama Minami, Tottori 680-8553, Japan

Moredun



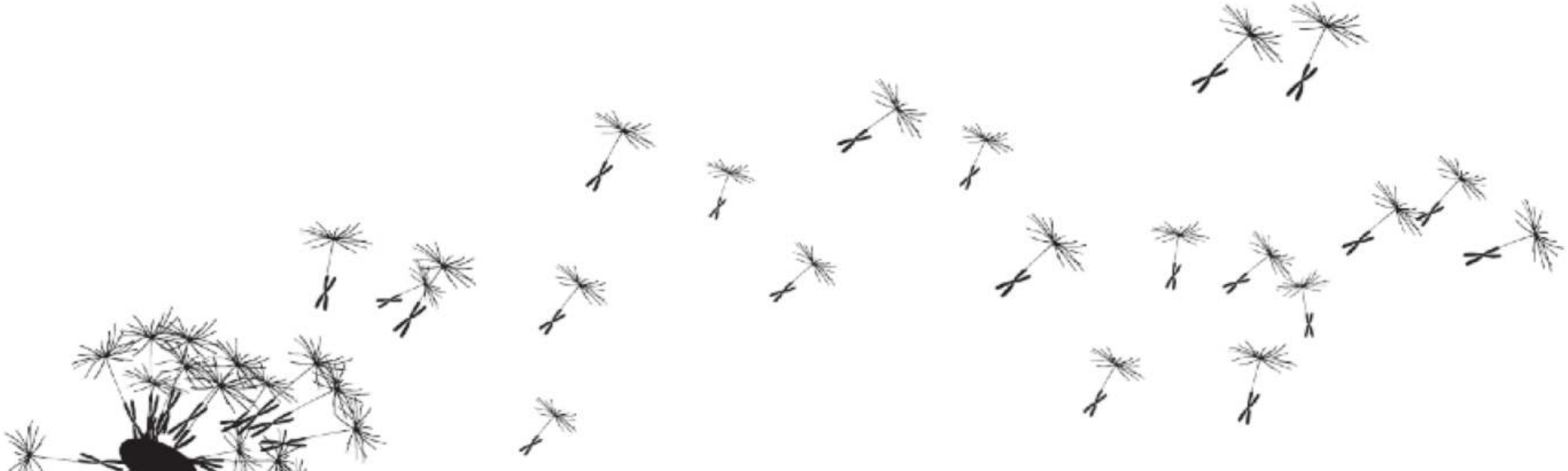
Draft Genome Assembly of the Poultry Red Mite, *Dermanyssus gallinae*

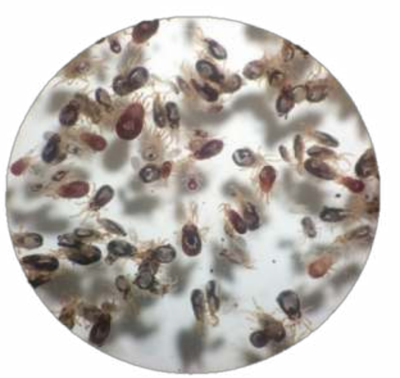
Stewart T. G. Burgess,[a] Kathryn Bartley,[a] Francesca Nunn,[a] Harry W. Wright,[a] Margaret Hughes,[b] Matthew Gemmell,[b] Sam Haldenby,[b] Steve Paterson,[b] Stephane Rombauts,[c,d,e] Fiona M. Tomley,[f] Damer P. Blake,[f] James Pritchard,[f] Sabine Schicht,[g] Christina Strube,[g] Øivind Øines,[h] Thomas Van Leeuwen,[i] Yves Van de Peer,[c,d,e,j] Alasdair J. Nisbet[a]

[a] Moredun Research Institute (MRI), Edinburgh, United Kingdom
[b] Centre for Genomic Research, Institute of Integrative Biology, University of Liverpool, Liverpool, United Kingdom
[c] Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium
[d] VIB Center for Plant Systems Biology, Ghent, Belgium
[e] Bioinformatics Institute Ghent, Ghent University, Ghent, Belgium
[f] Department of Pathology and Population Sciences, Royal Veterinary College, Hatfield, Hertfordshire, United Kingdom
[g] Institute for Parasitology, Centre for Infection Medicine, University of Veterinary Medicine Hannover, Hannover, Germany
[h] Norwegian Veterinary Institute, Oslo, Norway
[i] Department of Plants and Crops, Ghent University, Ghent, Belgium
[j] Department of Biochemistry, Genetics, and Microbiology, University of Pretoria, Pretoria, South Africa

Population genetics

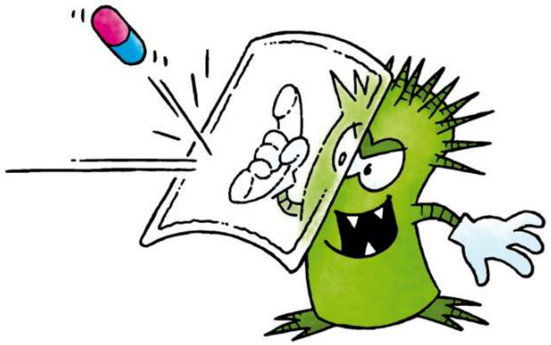
Population genetics is the study of genetic variation within and among populations and the evolutionary factors that explain this variation.





Population structure and diversity

Vaccination



Acaricide/drug resistance

Overall plan

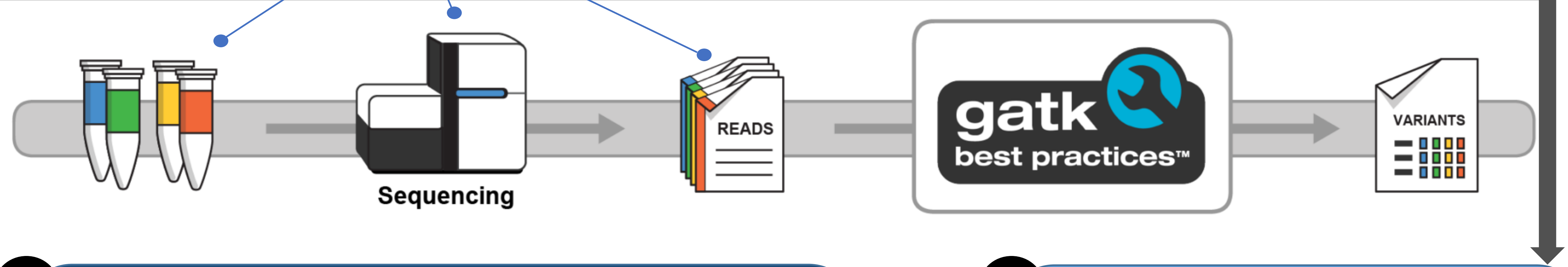
1

Hannover transcriptome – 454 sequencing[1]
Moredun genome assembly used as reference genome

2

Genome Analysis Toolkit

- GATK best practices followed for Germline SNP & Indel Discovery in Whole Genome and Exome



3

Variant calling file

- VCF stands for Variant Call Format.
- It is a standardized text file format for representing SNP, indel, and structural variation calls
- Used to identify a subset of SNPs with the highest confidence

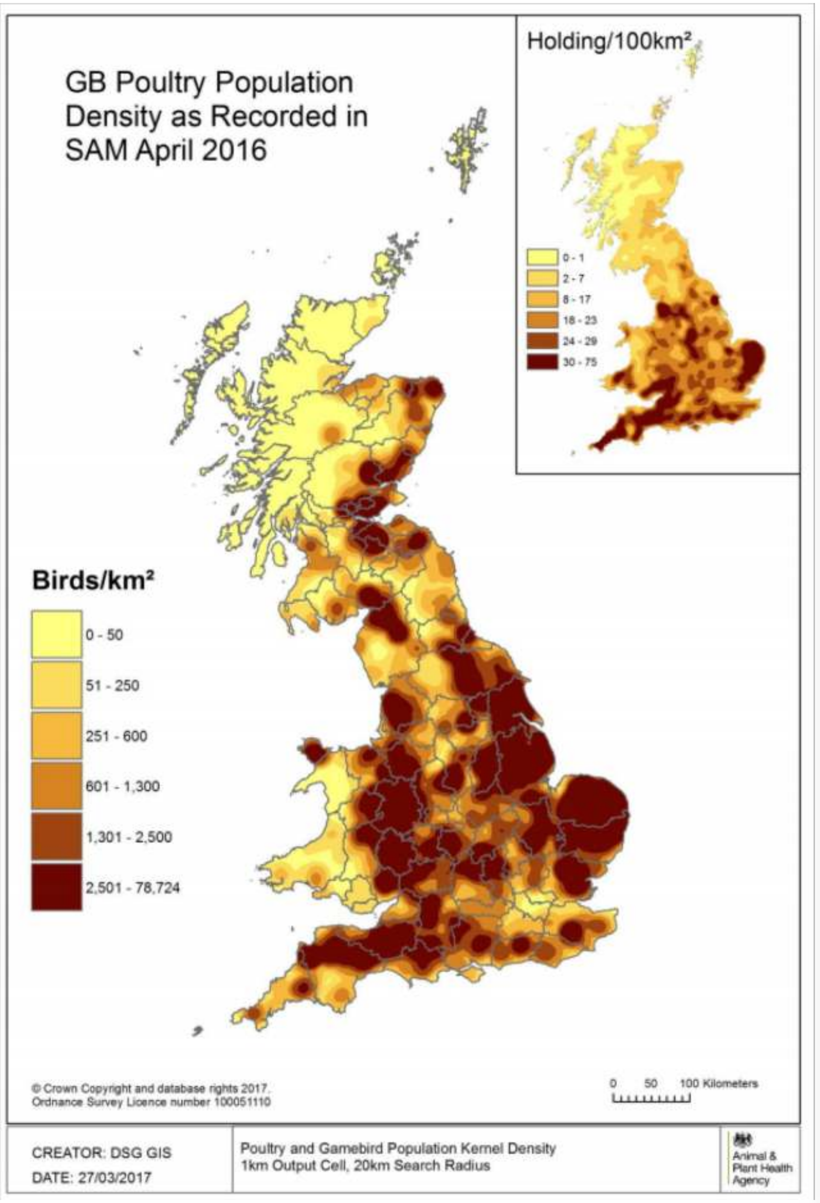
4

MassARRAY Panel

- MassARRAY panel designed to capacitate 384 samples per plate with up to 40 SNPs per sample
- Option to run two plates - 768 samples with 40 SNPs
- Mites from the UK and Europe
- Single mite samples

[1] Schict, S., Qi, W., Poveda, L. and Strube, C. (2014) 'Whole transcriptome analysis of the poultry red mite *Dermanyssus gallinae* (De Geer, 1778)', *Parasitology,* 141(3), pp. 336-46.

Sample collection: UK



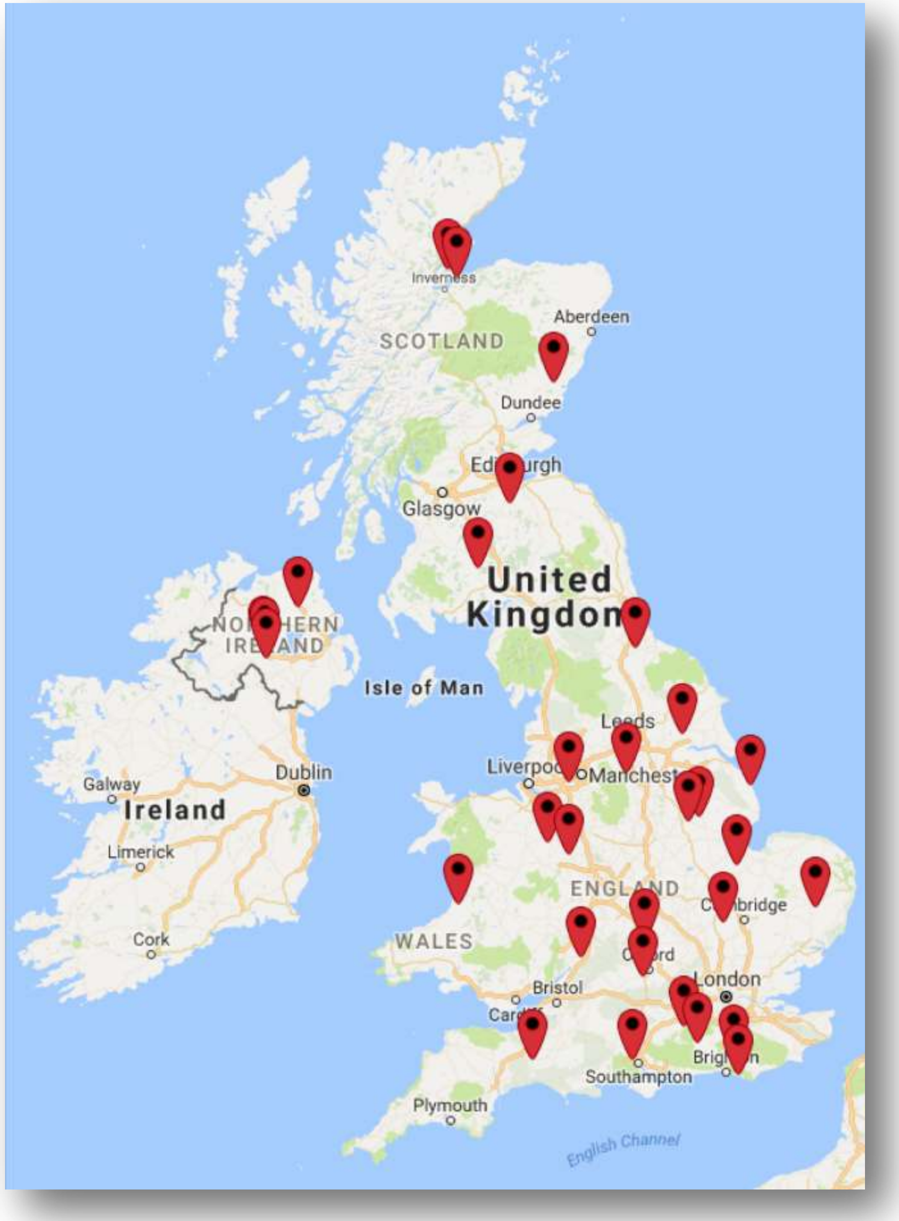
GB Poultry Population Density as Recorded in SAM April 2016

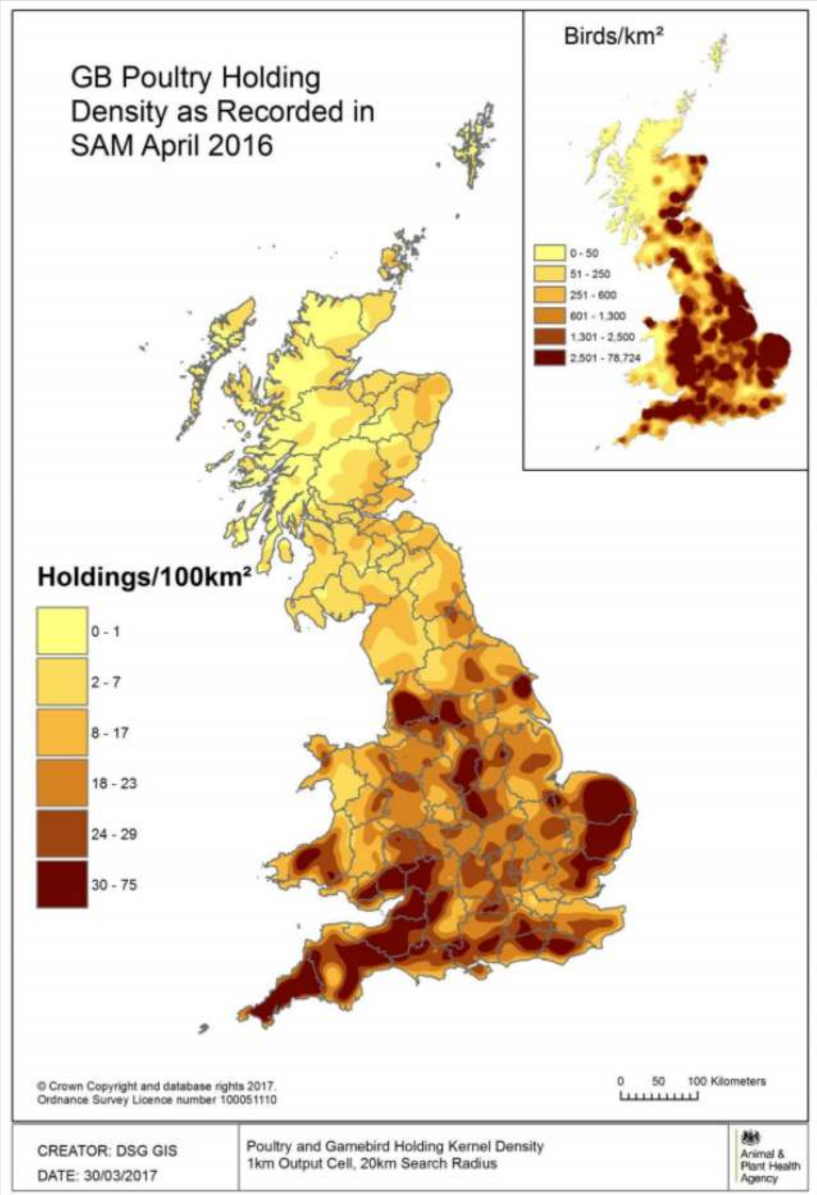
© Crown Copyright and database rights 2017.
Ordnance Survey Licence number 100051110

CREATOR: DSG GIS
DATE: 27/03/2017

Poultry and Gamebird Population Kernel Density
1km Output Cell, 20km Search Radius

Animal & Plant Health Agency





GB Poultry Holding Density as Recorded in SAM April 2016

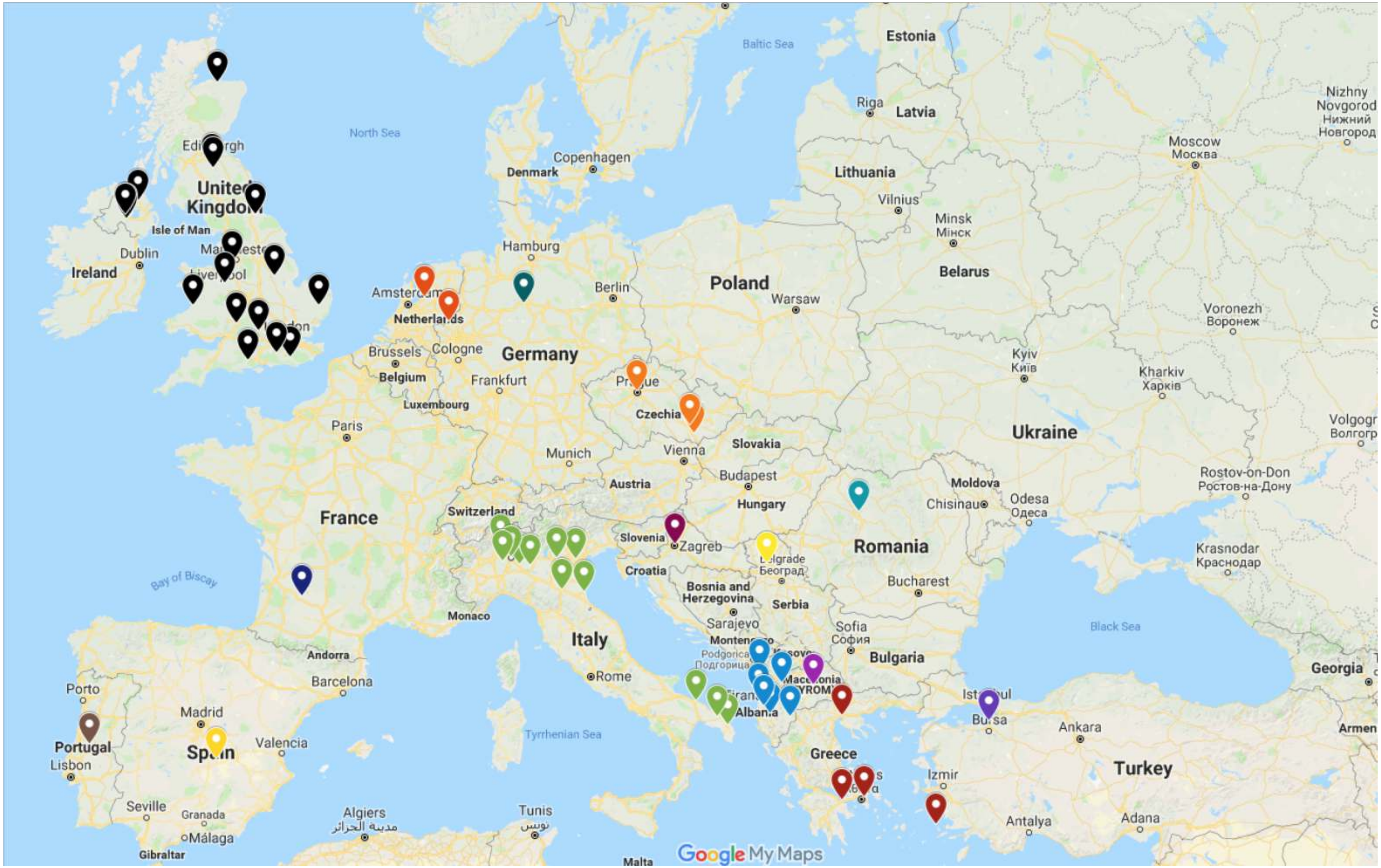
© Crown Copyright and database rights 2017.
Ordnance Survey Licence number 100051110

CREATOR: DSG GIS
DATE: 30/03/2017

Poultry and Gamebird Holding Kernel Density
1km Output Cell, 20km Search Radius

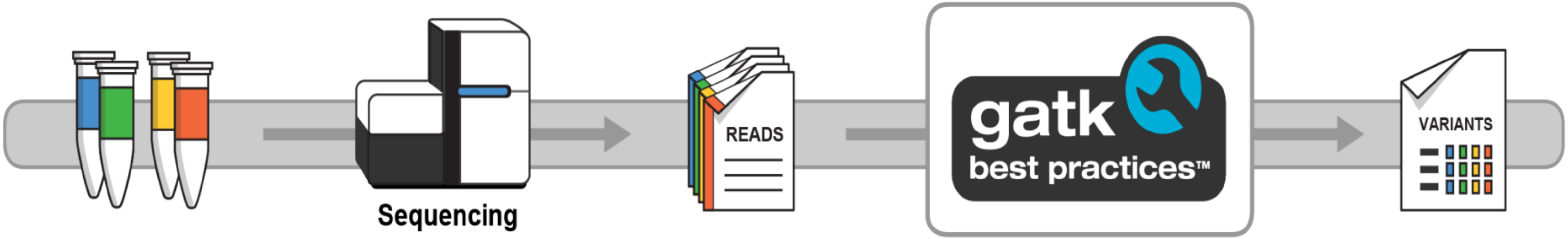
Animal & Plant Health Agency

Sample collection: Europe



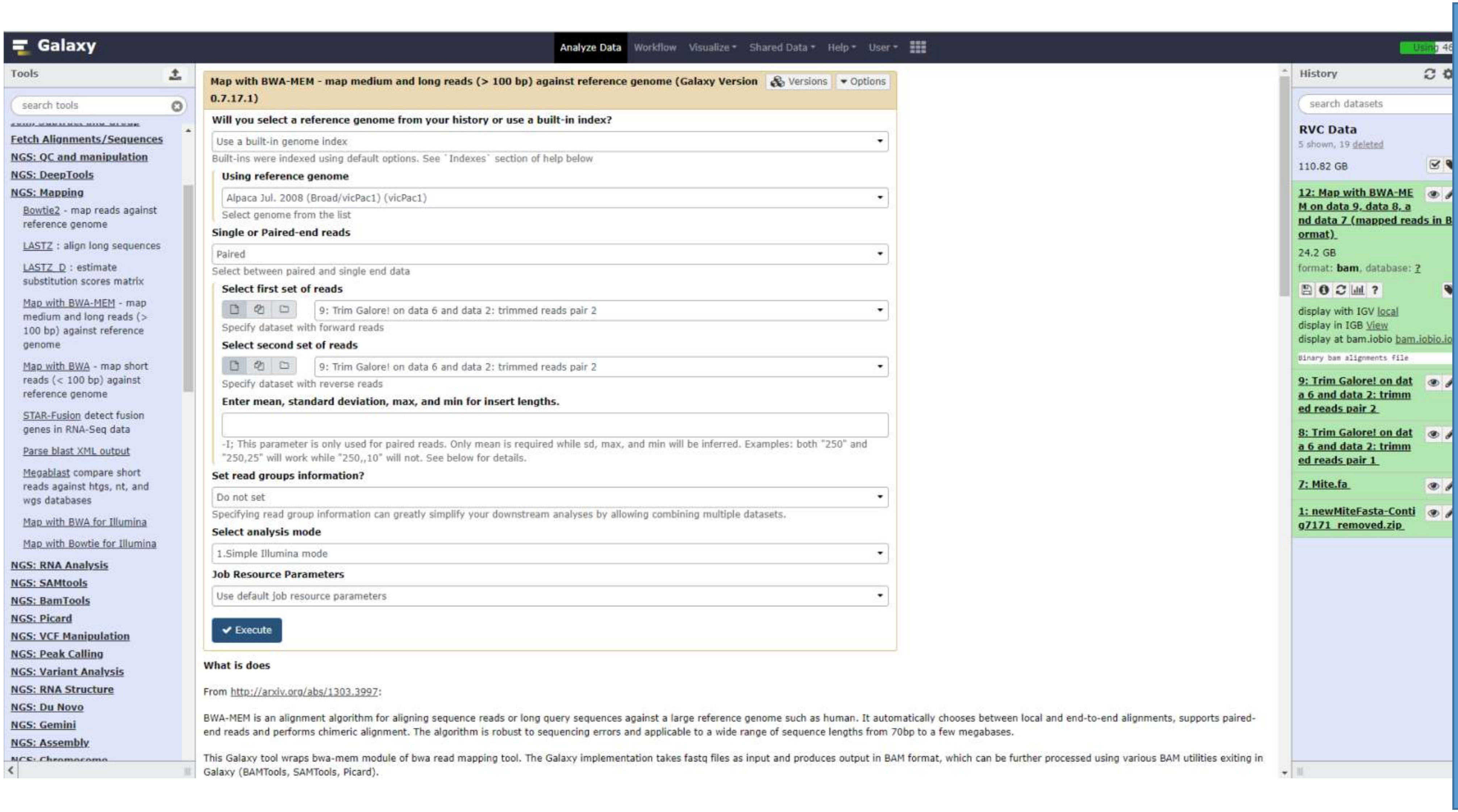
Genome Analysis Toolkit (GATK):

- Developed in the Data Sciences Platform at the Broad Institute, the toolkit offers a wide variety of tools with a primary focus on variant discovery and genotyping. Its powerful processing engine and high-performance computing features make it capable of taking on projects of any size.



GATK on Galaxy

- Galaxy servers implement a common core set of tools and reference genomes, and are open to anyone to use. They also contain tools and genomes that are local to each server.



Fetch Alignments/Sequences
NGS: QC and manipulation
NGS: DeepTools
NGS: Mapping
Bowtie2 - map reads against reference genome
LASTZ : align long sequences
LASTZ_D : estimate substitution scores matrix
Map with BWA-MEM - map medium and long reads (> 100 bp) against reference genome
Map with BWA - map short reads (< 100 bp) against reference genome
STAR-Fusion detect fusion genes in RNA-Seq data
Parse blast XML output
Megablast compare short reads against htgs, nt, and wgs databases
Map with BWA for Illumina
Map with Bowtie for Illumina
NGS: RNA Analysis
NGS: SAMtools
NGS: BamTools
NGS: Picard
NGS: VCF Manipulation
NGS: Peak Calling
NGS: Variant Analysis
NGS: RNA Structure
NGS: Du Novo
NGS: Gemini
NGS: Assembly

Map with BWA-MEM - map medium and long reads (> 100 bp) against reference genome (Galaxy Version 0.7.17.1) Versions Options

Will you select a reference genome from your history or use a built-in index?
Use a built-in genome index
Built-ins were indexed using default options. See `Indexes` section of help below

Using reference genome
Alpaca Jul. 2008 (Broad/vicPac1) (vicPac1)
Select genome from the list

Single or Paired-end reads
Paired
Select between paired and single end data

Select first set of reads
9: Trim Galore! on data 6 and data 2: trimmed reads pair 2
Specify dataset with forward reads

Select second set of reads
9: Trim Galore! on data 6 and data 2: trimmed reads pair 2
Specify dataset with reverse reads

Enter mean, standard deviation, max, and min for insert lengths.
-I; This parameter is only used for paired reads. Only mean is required while sd, max, and min will be inferred. Examples: both "250" and "250,25" will work while "250,,10" will not. See below for details.

Set read groups information?
Do not set
Specifying read group information can greatly simplify your downstream analyses by allowing combining multiple datasets.

Select analysis mode
1.Simple Illumina mode

Job Resource Parameters
Use default job resource parameters

✔ Execute

What is does

From http://arxiv.org/abs/1303.3997:

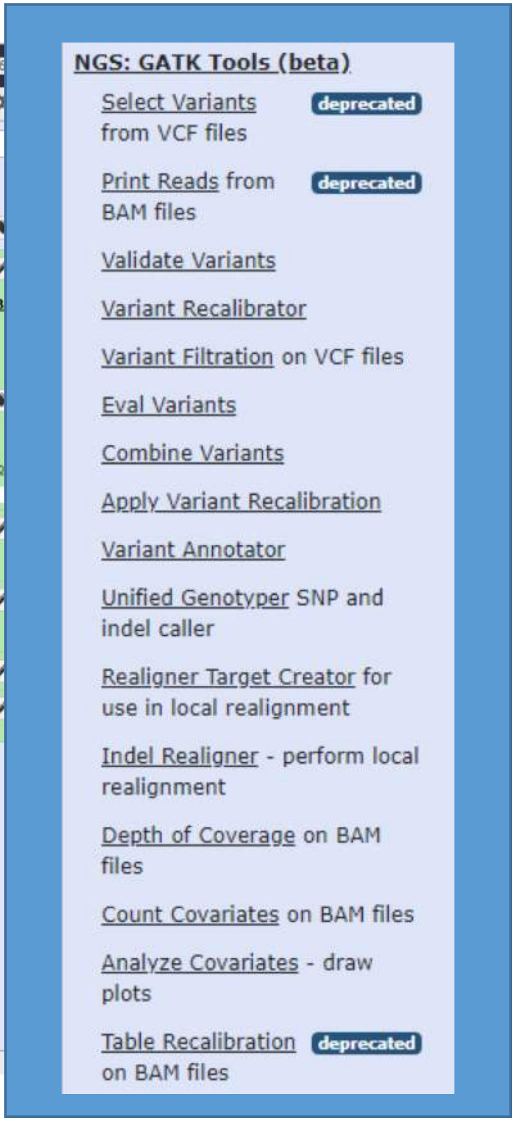
BWA-MEM is an alignment algorithm for aligning sequence reads or long query sequences against a large reference genome such as human. It automatically chooses between local and end-to-end alignments, supports paired-end reads and performs chimeric alignment. The algorithm is robust to sequencing errors and applicable to a wide range of sequence lengths from 70bp to a few megabases.

This Galaxy tool wraps bwa-mem module of bwa read mapping tool. The Galaxy implementation takes fastq files as input and produces output in BAM format, which can be further processed using various BAM utilities exiting in Galaxy (BAMTools, SAMTools, Picard).

RVC Data
5 shown, 19 deleted
110.82 GB

12: Map with BWA-MEM on data 9, data 8, and data 7 (mapped reads in BAM format)
24.2 GB
format: bam, database: ?
display with IGV local
display in IGB View
display at bam.iobio bam.iobio.io
Binary bam alignments file

9: Trim Galore! on data 6 and data 2: trimmed reads pair 2
8: Trim Galore! on data 6 and data 2: trimmed reads pair 1
7: Mite.fa
1: newMiteFasta-Contig7171_removed.zip

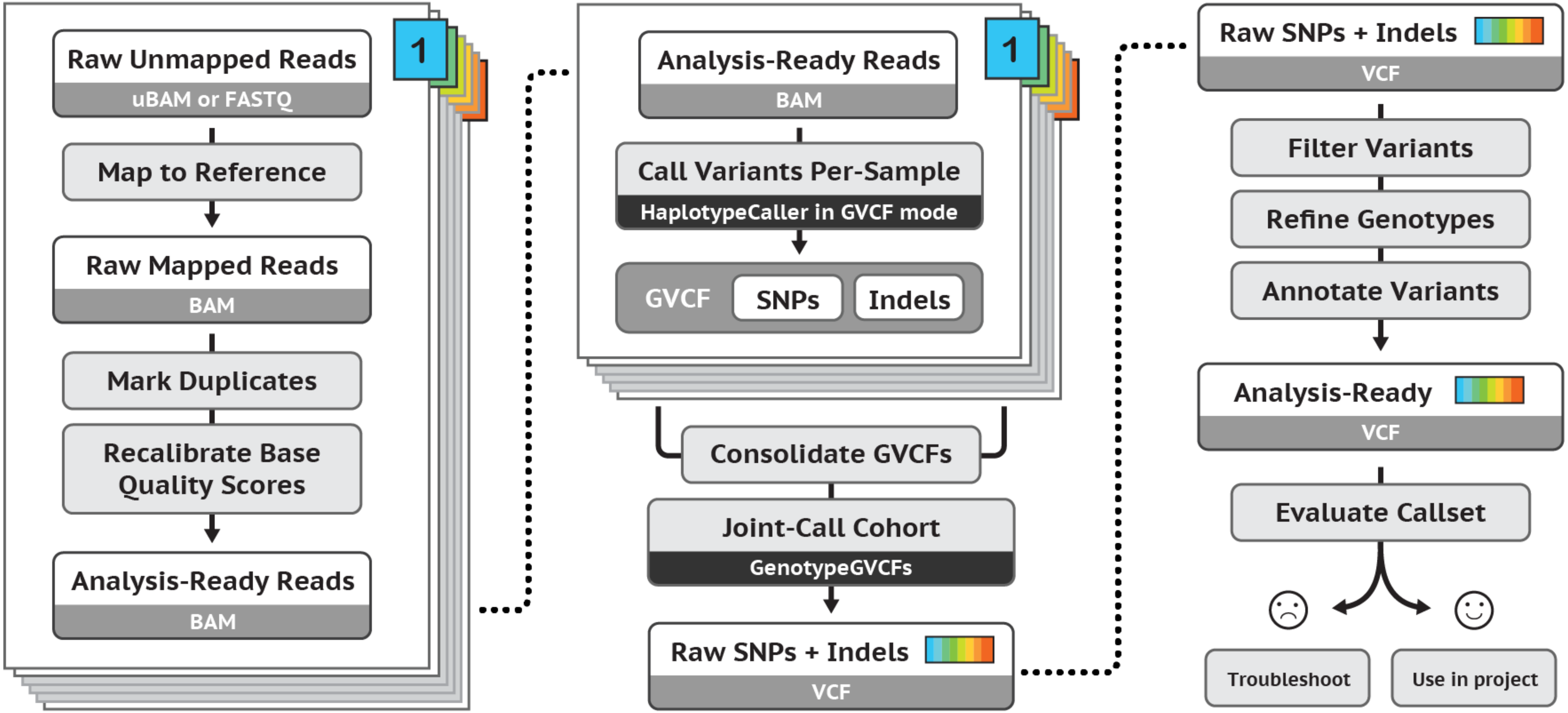


NGS: GATK Tools (beta)

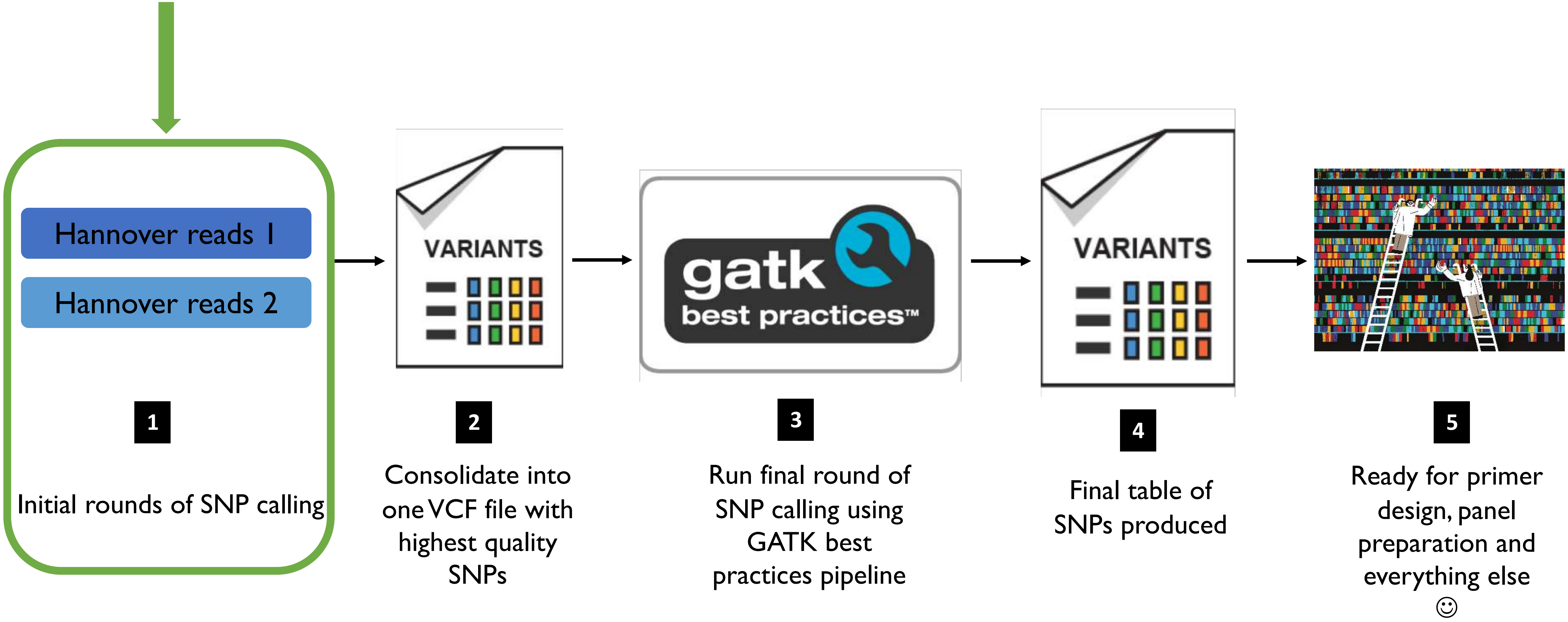
Select Variants from VCF files deprecated
Print Reads from BAM files deprecated
Validate Variants
Variant Recalibrator
Variant Filtration on VCF files
Eval Variants
Combine Variants
Apply Variant Recalibration
Variant Annotator
Unified Genotyper SNP and indel caller
Realigner Target Creator for use in local realignment
Indel Realigner - perform local realignment
Depth of Coverage on BAM files
Count Covariates on BAM files
Analyze Covariates - draw plots
Table Recalibration deprecated on BAM files

GATK's Best Practices

- Germline short variant discovery (SNPS + Indels)



GATK workflow



Hannover reads 1

Hannover reads 2

1

Initial rounds of SNP calling

2

Consolidate into one VCF file with highest quality SNPs

3

Run final round of SNP calling using GATK best practices pipeline

4

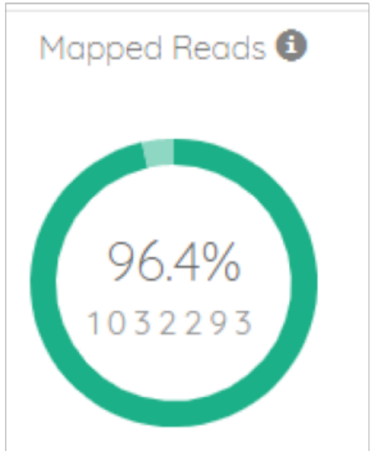
Final table of SNPs produced

5

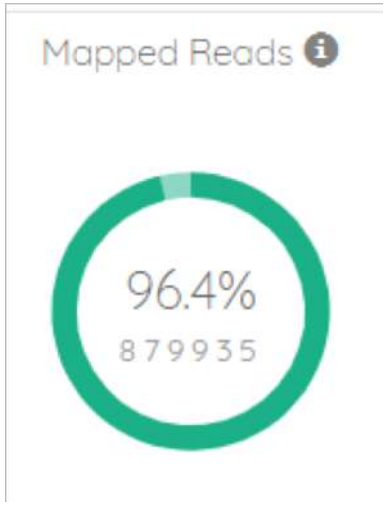
Ready for primer design, panel preparation and everything else ☺

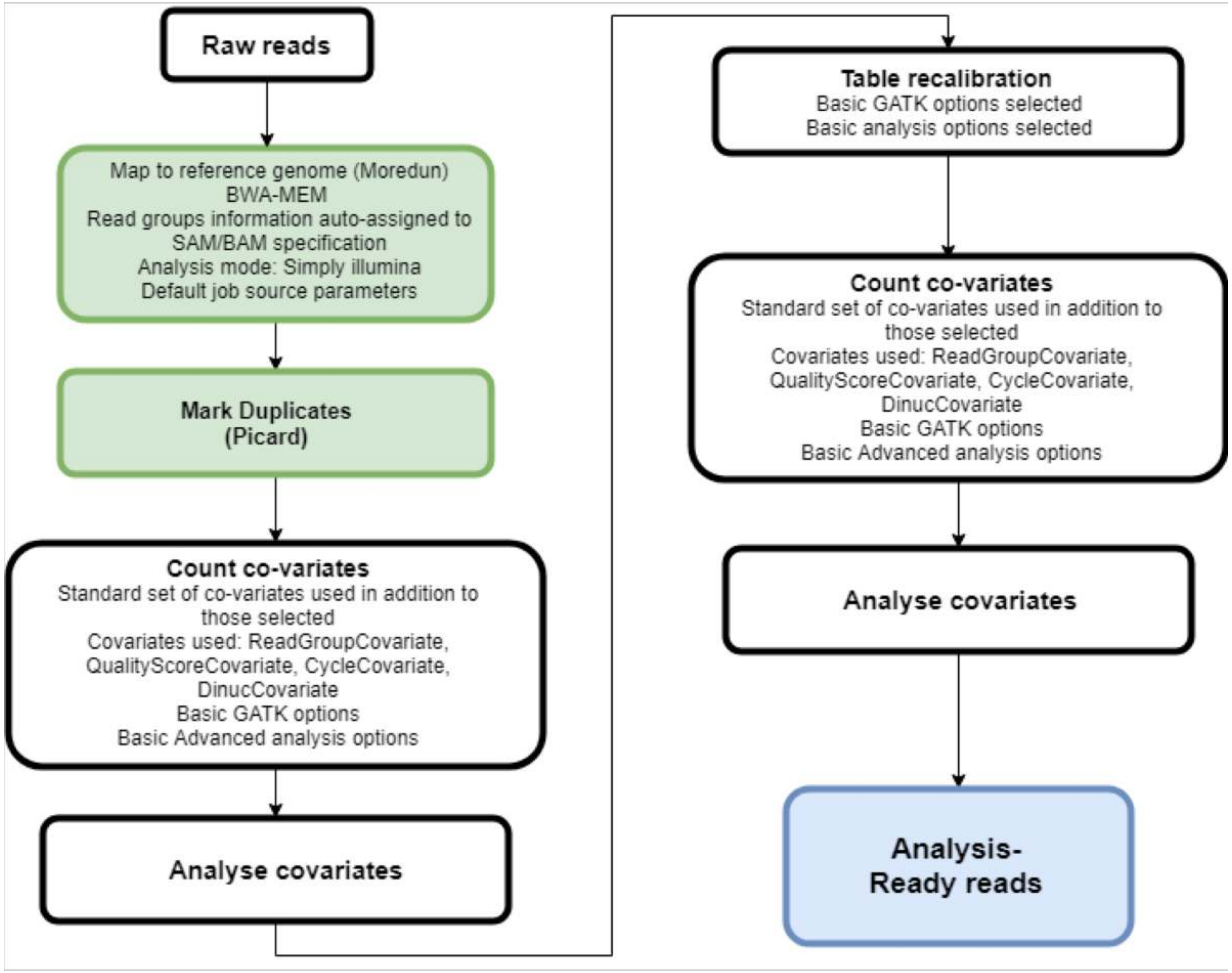
Pre-processing

Hannover 1



Hannover 2





Raw reads

Map to reference genome (Moredun)
BWA-MEM
Read groups information auto-assigned to SAM/BAM specification
Analysis mode: Simply illumina
Default job source parameters

Mark Duplicates (Picard)

Count co-variates
Standard set of co-variates used in addition to those selected
Covariates used: ReadGroupCovariate, QualityScoreCovariate, CycleCovariate, DinucCovariate
Basic GATK options
Basic Advanced analysis options

Analyse covariates

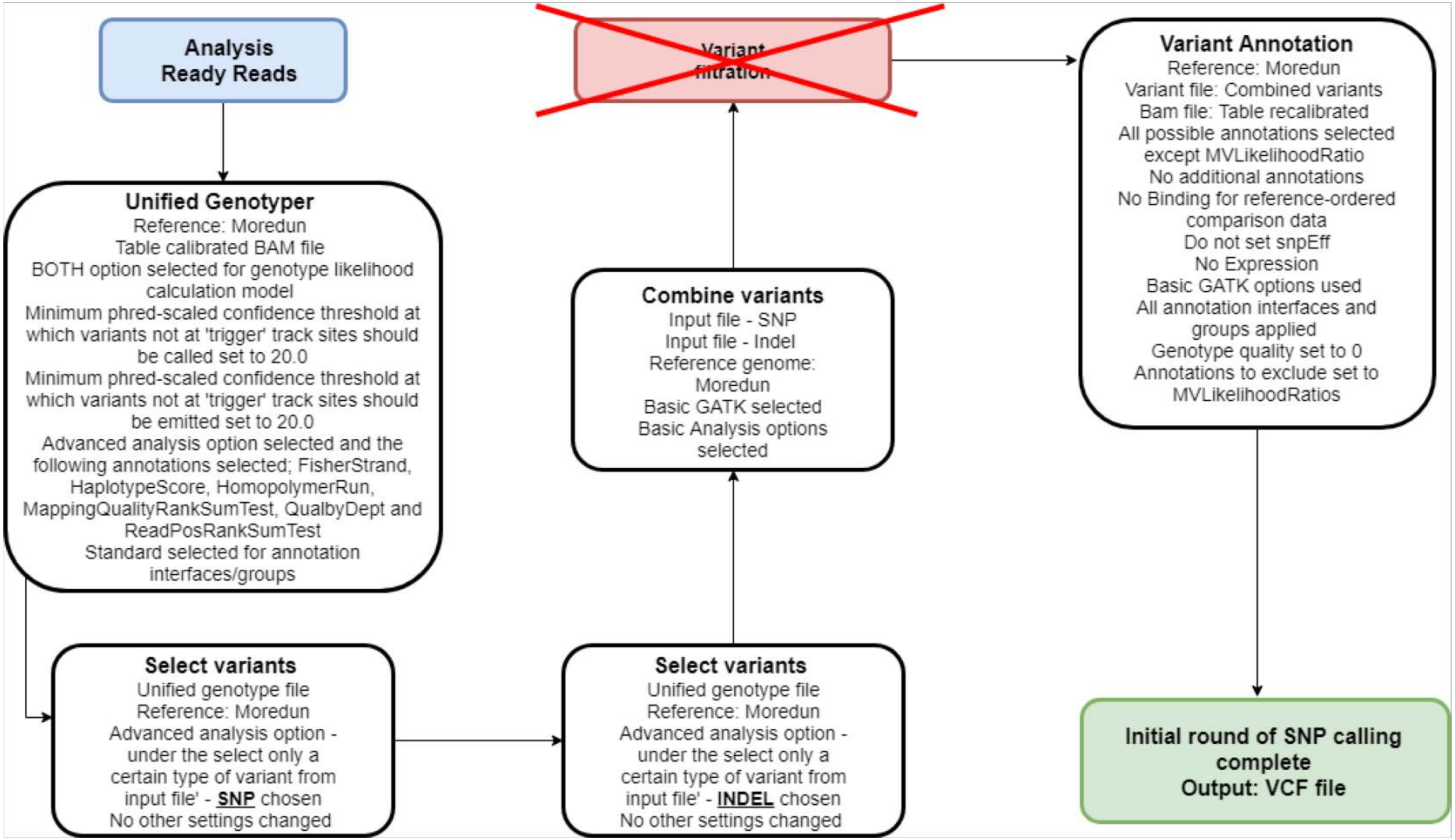
Table recalibration
Basic GATK options selected
Basic analysis options selected

Count co-variates
Standard set of co-variates used in addition to those selected
Covariates used: ReadGroupCovariate, QualityScoreCovariate, CycleCovariate, DinucCovariate
Basic GATK options
Basic Advanced analysis options

Analyse covariates

Analysis-Ready reads

Variant Discovery



Analysis Ready Reads

Unified Genotyper
Reference: Moredun
Table calibrated BAM file
BOTH option selected for genotype likelihood calculation model
Minimum phred-scaled confidence threshold at which variants not at 'trigger' track sites should be called set to 20.0
Minimum phred-scaled confidence threshold at which variants not at 'trigger' track sites should be emitted set to 20.0
Advanced analysis option selected and the following annotations selected; FisherStrand, HaplotypeScore, HomopolymerRun, MappingQualityRankSumTest, QualbyDept and ReadPosRankSumTest
Standard selected for annotation interfaces/groups

Select variants
Unified genotype file
Reference: Moredun
Advanced analysis option - under the select only a certain type of variant from input file' - **SNP** chosen
No other settings changed

Select variants
Unified genotype file
Reference: Moredun
Advanced analysis option - under the select only a certain type of variant from input file' - **INDEL** chosen
No other settings changed

Combine variants
Input file - SNP
Input file - Indel
Reference genome: Moredun
Basic GATK selected
Basic Analysis options selected

~~Variant filtration~~

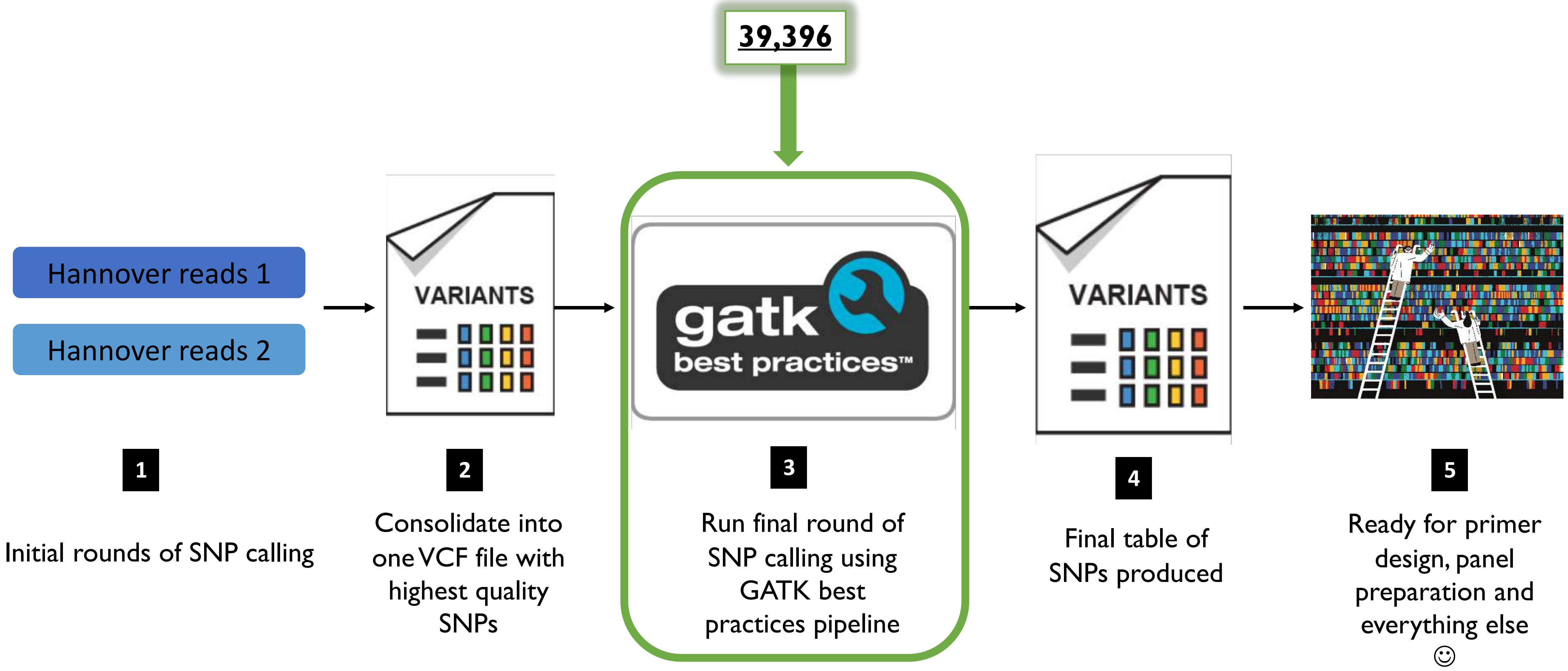
Variant Annotation
Reference: Moredun
Variant file: Combined variants
Bam file: Table recalibrated
All possible annotations selected except MVLikelihoodRatio
No additional annotations
No Binding for reference-ordered comparison data
Do not set snpEff
No Expression
Basic GATK options used
All annotation interfaces and groups applied
Genotype quality set to 0
Annotations to exclude set to MVLikelihoodRatios

Initial round of SNP calling complete
Output: VCF file

Initial SNP Tables

Dataset	Read sets	File size	Sequencing platform	Read type	Quality scores	GATK run1 complete	Mapping results	No. of SNPs	VCF intersect
Hannover	1	12.9MB	454	Single	Y	Y	96.4%	63,592	**39,396**
	2	13.5MB	454	Single	Y	Y	96.4%	69,440	

GATK workflow



1

Initial rounds of SNP calling

2

Consolidate into one VCF file with highest quality SNPs

3

Run final round of SNP calling using GATK best practices pipeline

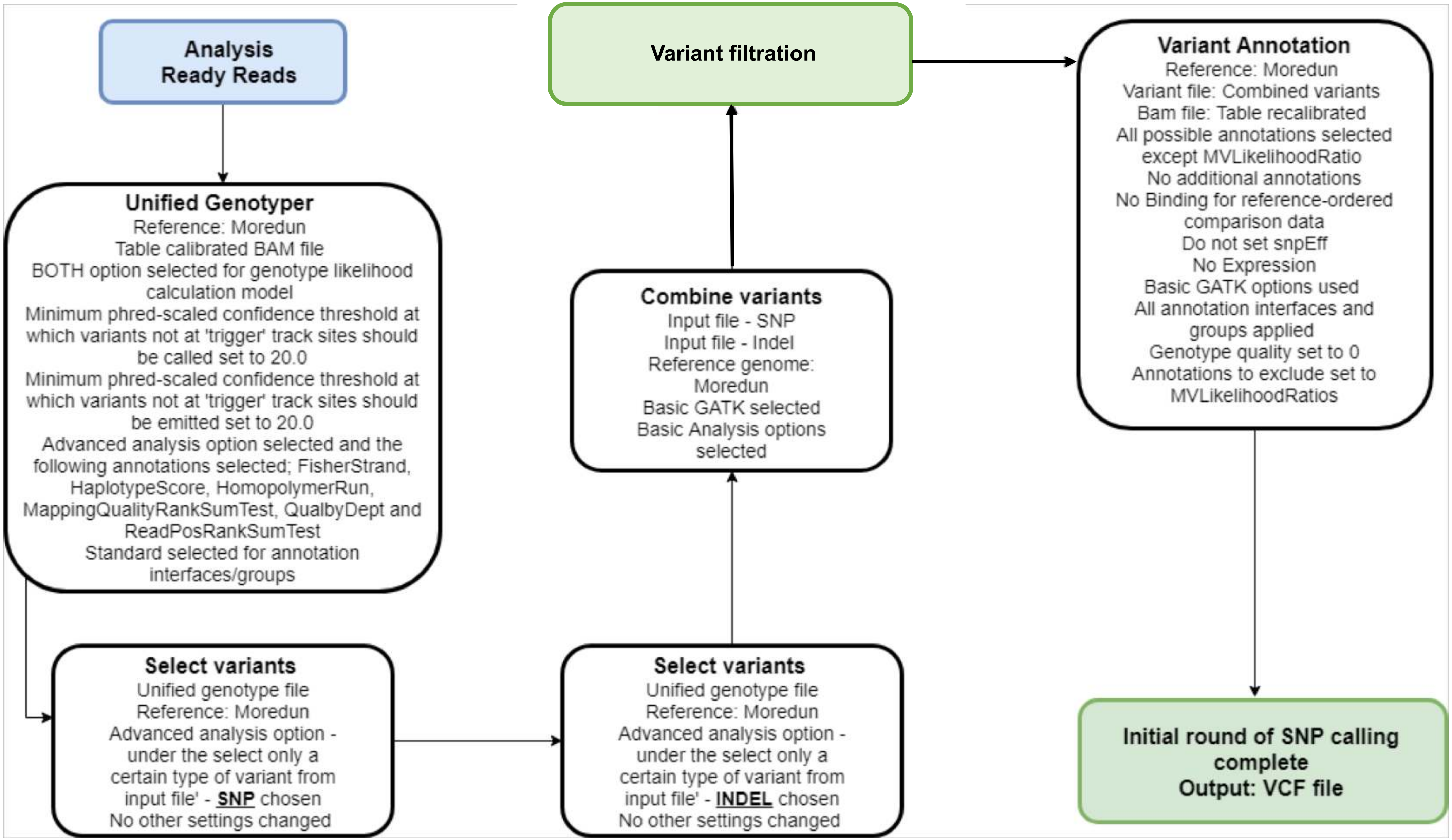
4

Final table of SNPs produced

5

Ready for primer design, panel preparation and everything else ☺

Variant Discovery



Analysis Ready Reads

Unified Genotyper
Reference: Moredun
Table calibrated BAM file
BOTH option selected for genotype likelihood calculation model
Minimum phred-scaled confidence threshold at which variants not at 'trigger' track sites should be called set to 20.0
Minimum phred-scaled confidence threshold at which variants not at 'trigger' track sites should be emitted set to 20.0
Advanced analysis option selected and the following annotations selected; FisherStrand, HaplotypeScore, HomopolymerRun, MappingQualityRankSumTest, QualbyDept and ReadPosRankSumTest
Standard selected for annotation interfaces/groups

Select variants
Unified genotype file
Reference: Moredun
Advanced analysis option - under the select only a certain type of variant from input file' - **SNP** chosen
No other settings changed

Select variants
Unified genotype file
Reference: Moredun
Advanced analysis option - under the select only a certain type of variant from input file' - **INDEL** chosen
No other settings changed

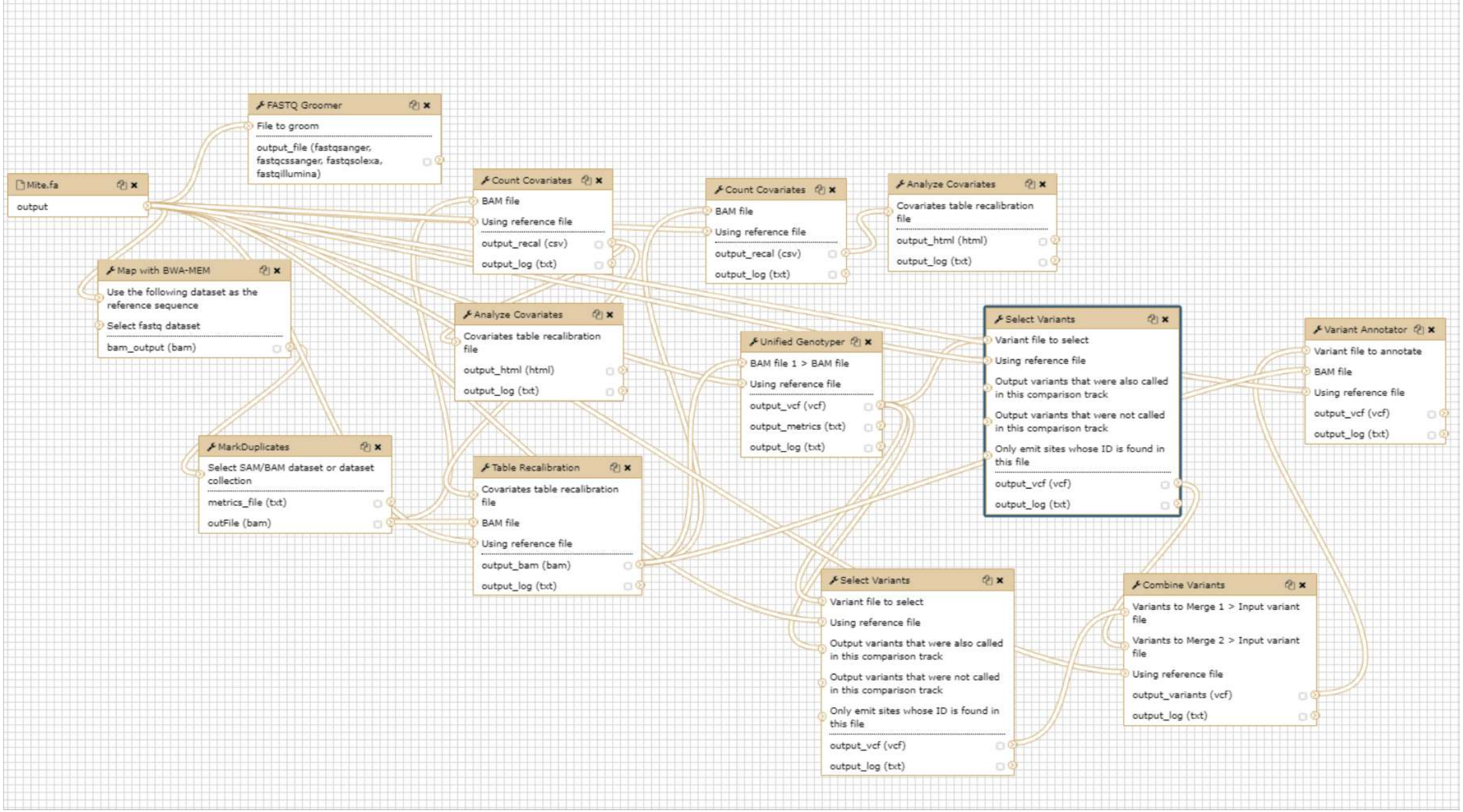
Combine variants
Input file - SNP
Input file - Indel
Reference genome: Moredun
Basic GATK selected
Basic Analysis options selected

Variant filtration

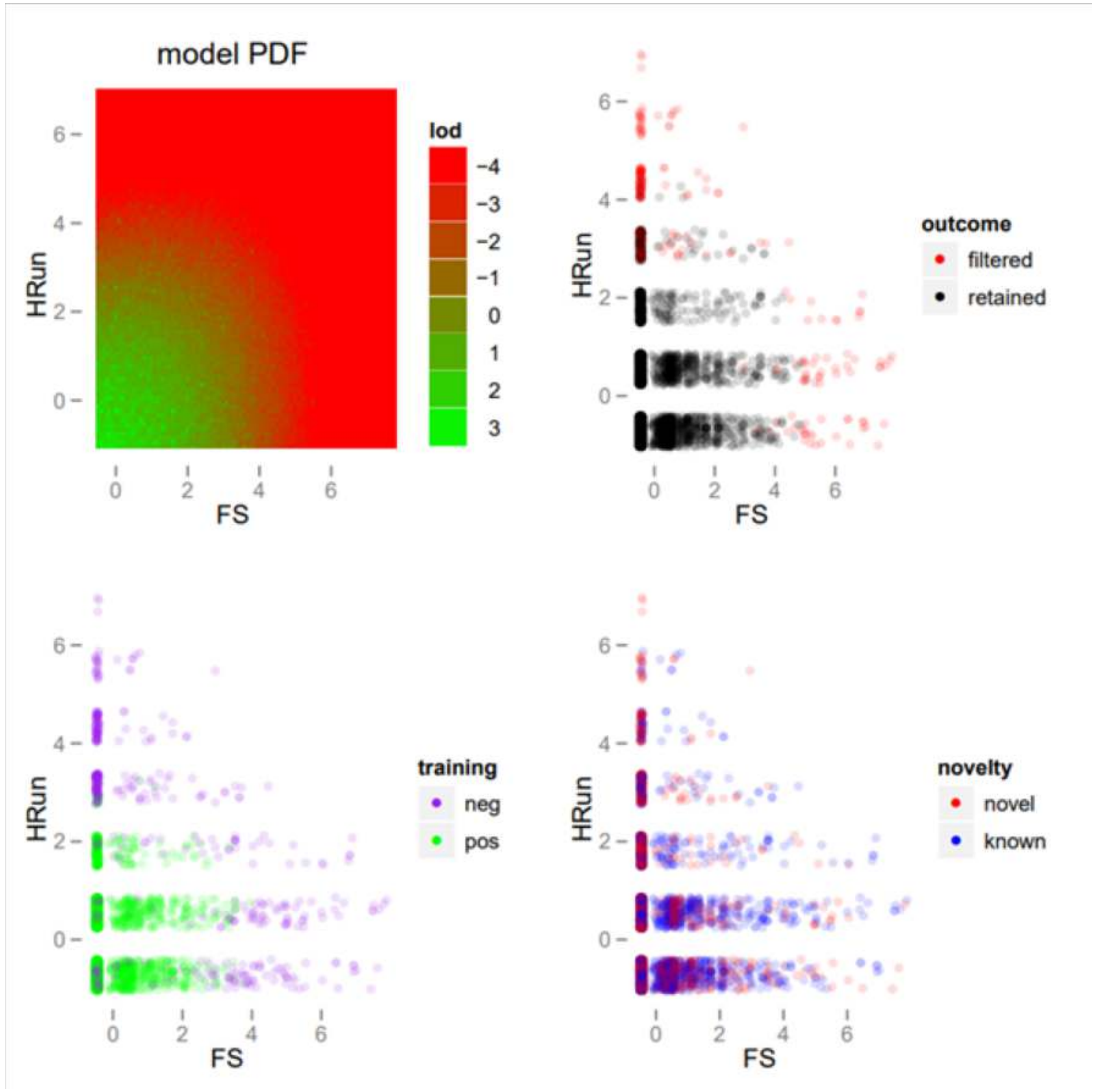
Variant Annotation
Reference: Moredun
Variant file: Combined variants
Bam file: Table recalibrated
All possible annotations selected except MVLikelihoodRatio
No additional annotations
No Binding for reference-ordered comparison data
Do not set snpEff
No Expression
Basic GATK options used
All annotation interfaces and groups applied
Genotype quality set to 0
Annotations to exclude set to MVLikelihoodRatios

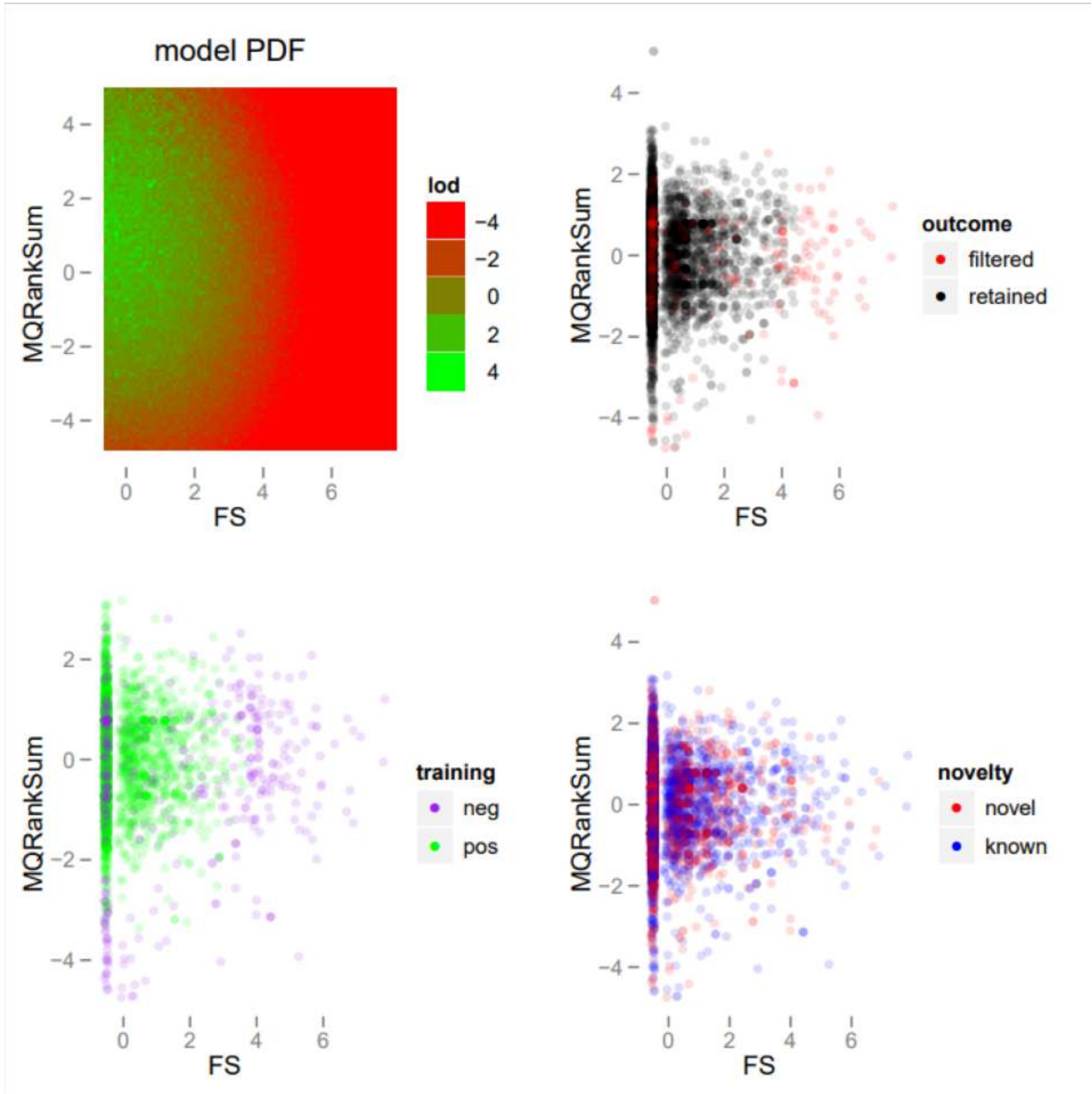
Initial round of SNP calling complete
Output: VCF file

In reality…


FASTQ Groomer
File to groom
output_file (fastqsanger, fastqcssanger, fastqsolexa, fastqillumina)
Mite.fa
output
Count Covariates
BAM file
Using reference file
output_recal (csv)
output_log (txt)
Count Covariates
BAM file
Using reference file
output_recal (csv)
output_log (txt)
Analyze Covariates
Covariates table recalibration file
output_html (html)
output_log (txt)
Map with BWA-MEM
Use the following dataset as the reference sequence
Select fastq dataset
bam_output (bam)
Analyze Covariates
Covariates table recalibration file
output_html (html)
output_log (txt)
Unified Genotyper
BAM file 1 > BAM file
Using reference file
output_vcf (vcf)
output_metrics (txt)
output_log (txt)
Select Variants
Variant file to select
Using reference file
Output variants that were also called in this comparison track
Output variants that were not called in this comparison track
Only emit sites whose ID is found in this file
output_vcf (vcf)
output_log (txt)
Variant Annotator
Variant file to annotate
BAM file
Using reference file
output_vcf (vcf)
output_log (txt)
MarkDuplicates
Select SAM/BAM dataset or dataset collection
metrics_file (txt)
outFile (bam)
Table Recalibration
Covariates table recalibration file
BAM file
Using reference file
output_bam (bam)
output_log (txt)
Select Variants
Variant file to select
Using reference file
Output variants that were also called in this comparison track
Output variants that were not called in this comparison track
Only emit sites whose ID is found in this file
output_vcf (vcf)
output_log (txt)
Combine Variants
Variants to Merge 1 > Input variant file
Variants to Merge 2 > Input variant file
Using reference file
output_variants (vcf)
output_log (txt)

Variant Filtration





Variant Filtration

Dataset	Read sets	GATK run1 complete	Mapping results	No. of SNPs	VCF intersect	GATK run 2 complete	No. of SNPs total	No. of SNPS PASS	No. of excluded SNPS	VCF intersect	VCF intersect PASS
Hannover	1	Y	96.4%	63,592	39,396	Y	66,296	65,248	1048	32,940	**<u>32,599</u>**
	2	Y	96.4%	69,440		Y	69,440	68294	1146		

-6,797 SNPs

Selecting SNPs

32,599 SNPs

1

High quality scoring SNPs in coding regions

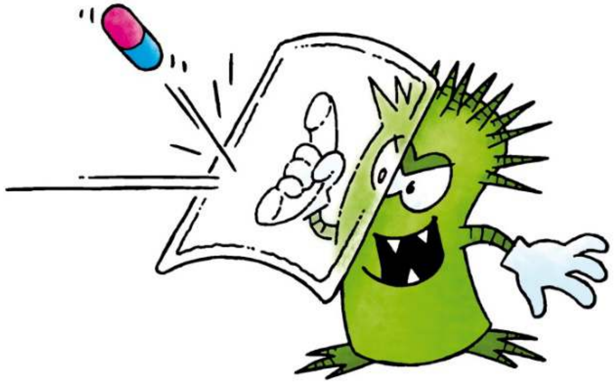
2

Bioinformatics tools to predict SNP effect

3

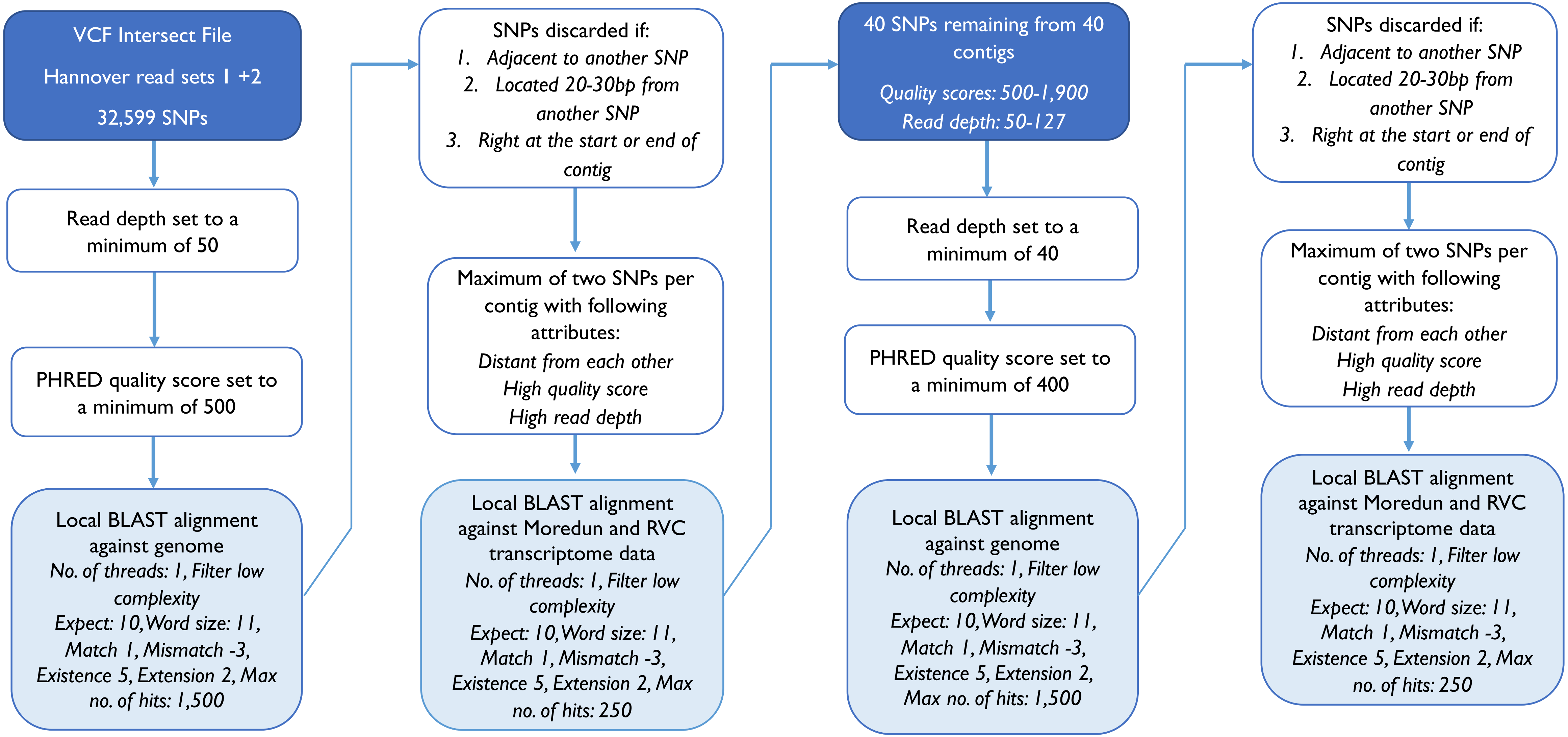
SNPs related to vaccine targets

4



SNPs relating to acaricide/drug resistance

Selecting SNPs



VCF Intersect File

Hannover read sets 1 +2

32,599 SNPs

Read depth set to a minimum of 50

PHRED quality score set to a minimum of 500

Local BLAST alignment against genome

No. of threads: 1, Filter low complexity

Expect: 10, Word size: 11, Match 1, Mismatch -3, Existence 5, Extension 2, Max no. of hits: 1,500

SNPs discarded if:

1. *Adjacent to another SNP*
2. *Located 20-30bp from another SNP*
3. *Right at the start or end of contig*

Maximum of two SNPs per contig with following attributes:

Distant from each other

High quality score

High read depth

Local BLAST alignment against Moredun and RVC transcriptome data

No. of threads: 1, Filter low complexity

Expect: 10, Word size: 11, Match 1, Mismatch -3, Existence 5, Extension 2, Max no. of hits: 250

40 SNPs remaining from 40 contigs

Quality scores: 500-1,900

Read depth: 50-127

Read depth set to a minimum of 40

PHRED quality score set to a minimum of 400

Local BLAST alignment against genome

No. of threads: 1, Filter low complexity

Expect: 10, Word size: 11, Match 1, Mismatch -3, Existence 5, Extension 2, Max no. of hits: 1,500

SNPs discarded if:

1. *Adjacent to another SNP*
2. *Located 20-30bp from another SNP*
3. *Right at the start or end of contig*

Maximum of two SNPs per contig with following attributes:

Distant from each other

High quality score

High read depth

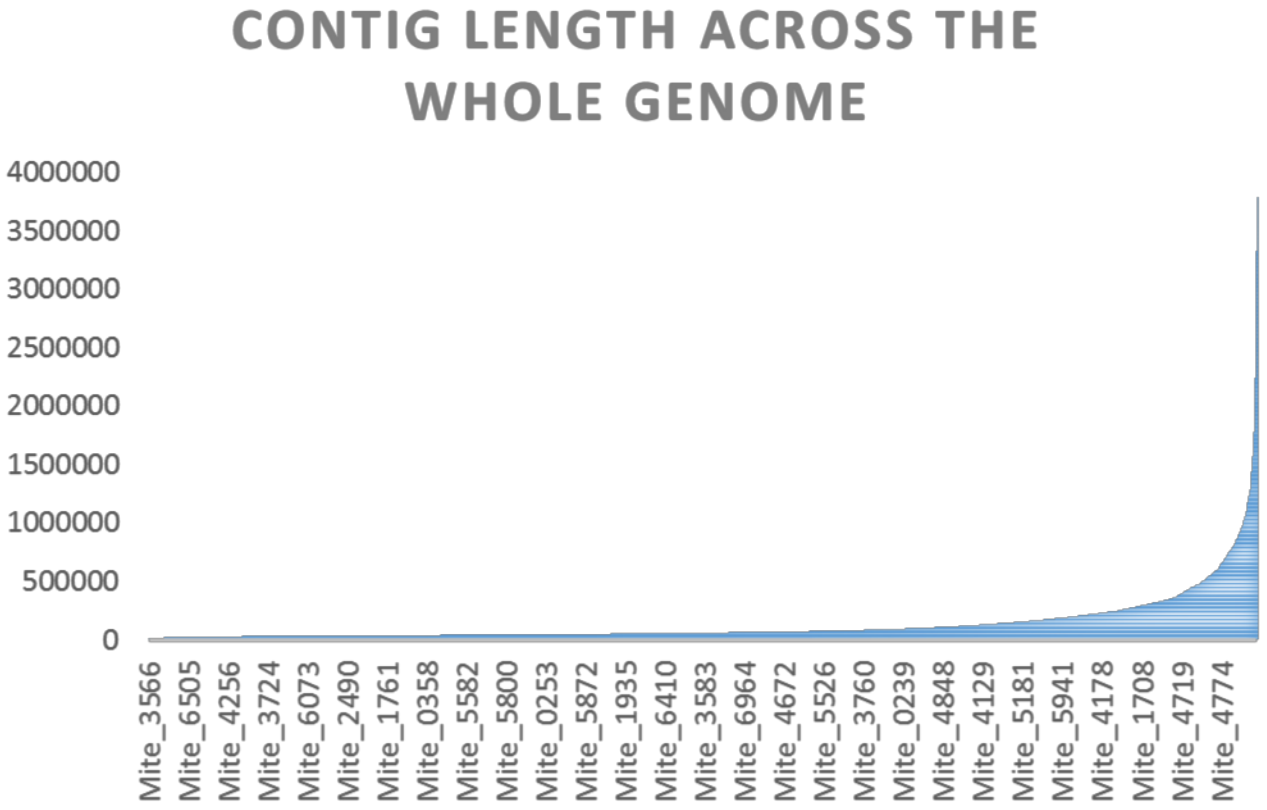
Local BLAST alignment against Moredun and RVC transcriptome data

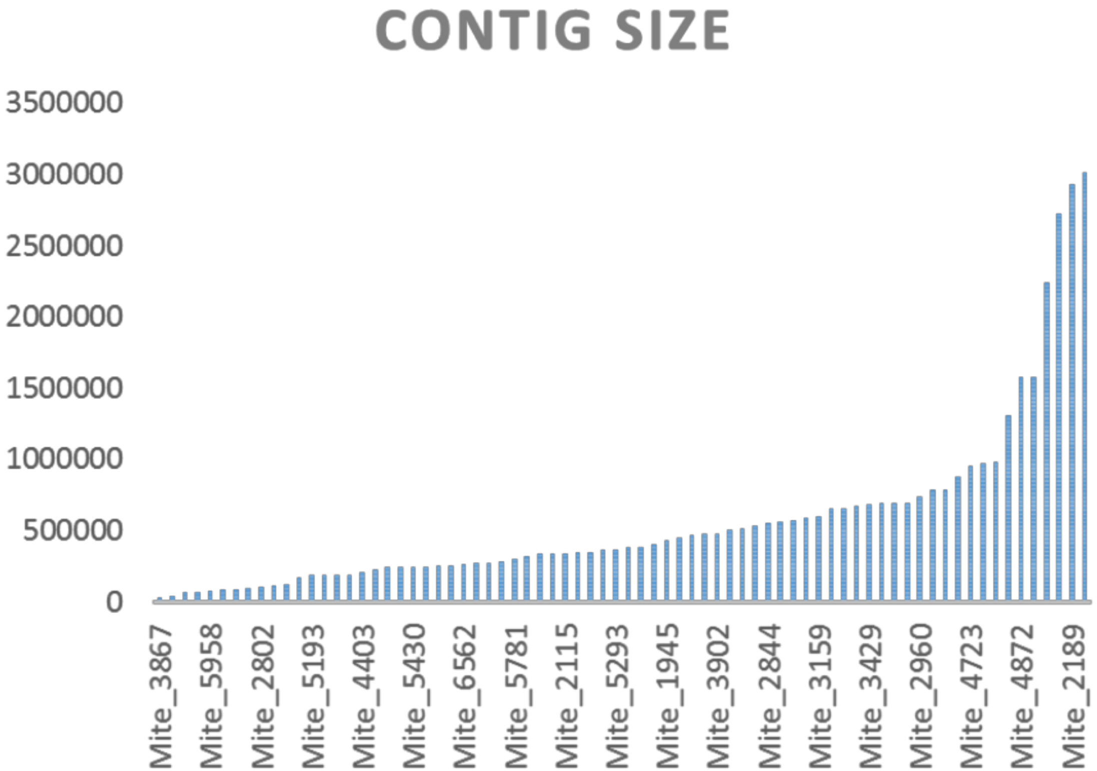
No. of threads: 1, Filter low complexity

Expect: 10, Word size: 11, Match 1, Mismatch -3, Existence 5, Extension 2, Max no. of hits: 250

Selecting SNPs

- 75 SNPS
- Contig size: 27,617-3,015,868
- Quality score: 400-2552
- Read depth: 40 - 127





What's next?

1

Validation of SNPs

- Selected SNPs tested in house
- Primer design followed by PCR and sequencing to further validate GATK workflow

2

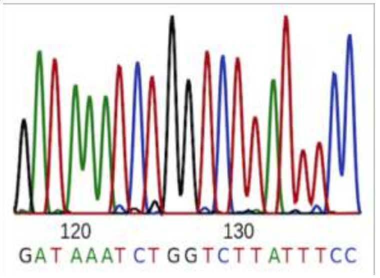
Send to MassARRAY

- Final panel of SNPS sent to MassARRAY company
- Primers designed to encompass target SNPs
- Sample preparation undertaken
- Wait for processing time

3

Computational analysis

- Diversity assessed across UK and European samples
- Inter-farm and intra-farm diversity investigated
- Phylogenetic analysis conducted



Acknowledgements

BEMB trust

RVC Royal Veterinary College University of London

Prof Damer Blake
Prof Fiona Tomley
Dr Dong Xia
Dr Tatiana Kuester
Miss Laura Evans

Moredun Research Institute

Dr Alasdair Nisbet
Dr Stewart Burgess
Dr Kath Bartley

Veterinærinstituttet Norwegian Veterinary Institute

Dr Øivind Øines

COREMI