Package: QUILT (via r-universe)

July 13, 2026

Type Package

Title QUILT

Version 2.0.4.9000

Date 2026-07-13

Description QUILT2 is an R and C++ program for fast genotype calling from sequencing reads using a large reference panel. QUILT2 is accurate and versatile, able to handle imputation from short read, long read, ancient DNA and cell-free DNA from NIPT.

Installation To install, first install dependencies, then run the install.packages command, pointing to the downloaded tarball (QUILT.tar.gz)

Getting started Please see github README.

Depends R (>= 4.0.0), STITCH (>= 1.8.2), mspbwt (>= 0.1.0), data.table

Remotes rwdavies/STITCH/STITCH, rwdavies/mspbwt/mspbwt

Imports Rcpp

LinkingTo Rcpp, RcppArmadillo, RcppEigen (>= 0.3.4.0.0)

RoxygenNote 7.3.2

License GPL (>= 3) | file LICENSE

SystemRequirements C++17

NeedsCompilation yes

Suggests testthat (>= 3.0.0)

Config/testthat/edition 3

Additional_repositories https://abiomix.r-universe.dev

URL https://github.com/abiomix/abiomix-r-open, https://github.com/rwdavies/QUILT

BugReports https://github.com/abiomix/abiomix-r-open/issues

Config/abiomix/public-license GPL-3.0-or-later

Config/abiomix/commercial-license Separate written agreement for Abiomix-owned contributions

Config/abiomix/license-policy https://github.com/abiomix/abiomix-r-open/blob/main/LICENSES.md

Repository https://abiomix.r-universe.dev

Date/Publication 2026-07-13 08:37:42 UTC

RemoteUrl https://github.com/abiomix/abiomix-r-open

RemoteRef HEAD

RemoteSha dd82290ba296385fcc7d7dfed339886e9deb41a2

RemoteSubdir packages/QUILT

Contents

`QUILT` *QUILT*

Description

QUILT

Usage

```
QUILT(
  outputdir,
  chr,
  method = "diploid",
  regionStart = NA,
  regionEnd = NA,
  buffer = NA,
  fflist = "",
  bamlist = "",
  cramlist = "",
  sampleNames_file = "",
  reference = "",
  nCores = 1,
  nGibbsSamples = 7,
  n_seek_its = 3,
  n_burn_in_seek_its = NA,
  Ksubset = 600,
  Knew = 600,
  K_top_matches = 5,
```

```
output_gt_phased_genotypes = TRUE,
heuristic_match_thin = 0.1,
output_filename = NULL,
RData_objects_to_save = NULL,
output_RData_filename = NULL,
prepared_reference_filename = "",
save_prepared_reference = FALSE,
tempdir = NA,
bqFilter = as.integer(17),
useSoftClippedBases = FALSE,
panel_size = NA,
posfile = "",
genfile = "",
phasefile = "",
maxDifferenceBetweenReads = 1e+10,
make_plots = FALSE,
make_plots_block_gibbs = FALSE,
verbose = TRUE,
shuffle_bin_radius = 5000,
iSizeUpperLimit = 1e+06,
output_read_label_prob = FALSE,
record_read_label_usage = FALSE,
record_interim_dosages = FALSE,
use_bx_tag = TRUE,
bxTagUpperLimit = 50000,
addOptimalHapsToVCF = FALSE,
estimate_bq_using_truth_read_labels = FALSE,
override_default_params_for_small_ref_panel = TRUE,
gamma_physically_closest_to = NA,
seed = NA,
hla_run = FALSE,
downsampleToCov = 30,
minGLValue = 1e-10,
minimum_number_of_sample_reads = 2,
nGen = NA,
reference_vcf_file = "",
reference_haplotype_file = "",
reference_legend_file = "",
reference_sample_file = "",
reference_populations = NA,
reference_phred = 30,
reference_exclude_samplelist_file = "",
region_exclude_file = "",
genetic_map_file = "",
nMaxDH = NA,
make_fake_vcf_with_sites_list = FALSE,
output_sites_filename = NA,
expRate = 1,
```

```
  maxRate = 100,
  minRate = 0.1,
  print_extra_timing_information = FALSE,
  small_ref_panel_block_gibbs_iterations = c(3, 6, 9),
  small_ref_panel_gibbs_iterations = 20,
  plot_per_sample_likelihoods = FALSE,
  use_small_eHapsCurrent_tc = FALSE,
  mspbwtL = 3,
  mspbwtM = 1,
  use_mspbwt = FALSE,
  mspbwt_nindices = 4L,
  use_splitreadgl = FALSE,
  override_use_eMatDH_special_symbols = NA,
  use_hapMatcherR = TRUE,
  shard_check_every_pair = TRUE,
  use_eigen = TRUE,
  impute_rare_common = FALSE,
  rare_af_threshold = 0.001,
  make_heuristic_plot = FALSE,
  heuristic_approach = "A",
  use_list_of_columns_of_A = TRUE,
  calculate_gamma_on_the_fly = TRUE
)
```

Arguments

`outputdir`	What output directory to use
`chr`	What chromosome to run. Should match BAM headers
`method`	What method to run (diploid or nipt)
`regionStart`	When running imputation, where to start from. The 1-based position x is kept if regionStart <= x <= regionEnd
`regionEnd`	When running imputation, where to stop
`buffer`	Buffer of region to perform imputation over. So imputation is run form regionStart-buffer to regionEnd+buffer, and reported for regionStart to regionEnd, including the bases of regionStart and regionEnd
`fflist`	Path to file with fetal fraction values, one row per entry, in the same order as the bamlist
`bamlist`	Path to file with bam file locations. File is one row per entry, path to bam files. Bam index files should exist in same directory as for each bam, suffixed either .bam.bai or .bai
`cramlist`	Path to file with cram file locations. File is one row per entry, path to cram files. cram files are converted to bam files on the fly for parsing into QUILT
`sampleNames_file`	Optional, if not specified, sampleNames are taken from the SM tag in the header of the BAM / CRAM file. This argument is the path to file with sampleNames for samples. It is used directly to name samples in the order they appear in the bamlist / cramlist

`reference` Path to reference fasta used for making cram files. Only required if cramlist is defined

`nCores` How many cores to use

`nGibbsSamples` How many Gibbs samples to use

`n_seek_its` How many iterations between first using current haplotypes to update read labels, and using current read labels to get new reference haplotypes, to perform

`n_burn_in_seek_its`
How many iterations of the seek_its should be burn in. As an example, if n_seek_its is 3 and n_burn_in_seek_its is 2, then only the dosage from the final round is included. If n_seek_its is 4 and n_burn_in_seek_its is 2, then dosages from the last two rounds are used. Default value NA sets n_burn_in_seek_its to n_seek_its minus 1

`Ksubset` How many haplotypes to use in the faster Gibbs sampling

`Knew` How many haplotypes to replace per-iteration after doing the full reference panel imputation

`K_top_matches` How many top haplotypes to store in each grid site when looking for good matches in the full haplotype reference panel. Large values potentially bring in more haplotype diversity, but risk losing haplotypes that are good matches over shorter distances

`output_gt_phased_genotypes`
When TRUE, output GT entry contains phased genotypes (haplotypes). When FALSE, it is from the genotype posteriors, and masked when the maximum genotype posterior entry is less than 0.9

`heuristic_match_thin`
What fraction of grid sites to use when looking for good matches in the full haplotype reference panel. Smaller values run faster but potentially miss haplotypes

`output_filename`
Override the default bgzip-VCF / bgen output name with this given file name. Please note that this does not change the names of inputs or outputs (e.g. RData, plots), so if outputdir is unchanged and if multiple QUILT runs are processing on the same region then they may over-write each others inputs and outputs.

`RData_objects_to_save`
Can be used to name interim and misc results from imputation to save an an RData file. Default NULL means do not save such output

`output_RData_filename`
Override the default location for miscellaneous outputs saved in RData format

`prepared_reference_filename`
Optional path to prepared RData file with reference objects. Can be used instead of outputdir to coordinate use of QUILT_prepare_reference and QUILT

`save_prepared_reference`
If preparing reference as part of running QUILT, whether to save the prepared reference output file. Note that if the reference was already made using QUILT_prepare_reference, this is ignored

`tempdir` What directory to use as temporary directory. If set to NA, use default R tempdir. If possible, use ramdisk, like /dev/shm/

bqFilter Minimum BQ for a SNP in a read. Also, the algorithm uses bq<=mq, so if mapping quality is less than this, the read isnt used

useSoftClippedBases
Whether to use (TRUE) or not use (FALSE) bases in soft clipped portions of reads

panel_size Integer number of reference haplotypes to use, set to NA to use all of them

posfile Optional, only needed when using genfile or phasefile. File with positions of where to impute, lining up one-to-one with genfile. File is tab seperated with no header, one row per SNP, with col 1 = chromosome, col 2 = physical position (sorted from smallest to largest), col 3 = reference base, col 4 = alternate base. Bases are capitalized. Example first row: 1<tab>1000<tab>A<tab>G<tab>

genfile Path to gen file with high coverage results. Empty for no genfile. If both genfile and phasefile are given, only phasefile is used, as genfile (unphased genotypes) is derivative to phasefile (phased genotypes). File has a header row with a name for each sample, matching what is found in the bam file. Each subject is then a tab seperated column, with 0 = hom ref, 1 = het, 2 = hom alt and NA indicating missing genotype, with rows corresponding to rows of the posfile. Note therefore this file has one more row than posfile which has no header

phasefile Path to phase file with truth phasing results. Empty for no phasefile. Supercedes genfile if both options given. File has a header row with a name for each sample, matching what is found in the bam file. Each subject is then a tab seperated column, with 0 = ref and 1 = alt, separated by a vertical bar |, e.g. 0|0 or 0|1. Note therefore this file has one more row than posfile which has no header.

maxDifferenceBetweenReads
How much of a difference to allow the reads to make in the forward backward probability calculation. For example, if P(read | state 1)=1 and P(read | state 2)=1e-6, re-scale so that their ratio is this value. This helps prevent any individual read as having too much of an influence on state changes, helping prevent against influence by false positive SNPs

make_plots Whether to make some plots of per-sample imputation. Especially nice when truth data. This is pretty slow though so useful more for debugging and understanding and visualizing performance

make_plots_block_gibbs
Whether to make some plots of per-sample imputation looking at how the block Gibbs is performing. This can be extremely slow so use for debugging or visualizing performance on one-off situations not for general runs

verbose whether to be more verbose when running

shuffle_bin_radius
Parameter that controls how to detect ancestral haplotypes that are shuffled during EM for possible re-setting. If set (not NULL), then recombination rate is calculated around pairs of SNPs in window of twice this value, and those that exceed what should be the maximum (defined by nGen and maxRate) are checked for whether they are shuffled

iSizeUpperLimit
Do not use reads with an insert size of more than this value

output_read_label_prob
: Whether to output read labels after the final Gibbs samplings (i.e. whether reads were assigned to arbitrary labelled haplotype 1 or 2 and the probability)

record_read_label_usage
: Whether to store what read labels were used during the Gibbs samplings (i.e. whether reads were assigned to arbitrary labelled haplotype 1 or 2)

record_interim_dosages
: Whether to record interim dosages or not

use_bx_tag
: Whether to try and use BX tag in same to indicate that reads come from the same underlying molecule

bxTagUpperLimit
: When using BX tag, at what distance between reads to consider reads with the same BX tag to come from different molecules

addOptimalHapsToVCF
: Whether to add optimal haplotypes to vcf when phasing information is present, where optimal is imputation done when read label origin is known

estimate_bq_using_truth_read_labels
: When using phasefile with known truth haplotypes, infer truth read labels, and use them to infer the real base quality against the bam recorded base qualities

override_default_params_for_small_ref_panel
: When set to TRUE, then when using a smaller reference panel size (fewer haplotypes than Ksubset), parameter choices are reset appropriately. When set to FALSE, original values are used, which might crash QUILT

gamma_physically_closest_to
: For HLA imputation, the physical position closest to the centre of the gene

seed
: The seed that controls random number generation. When NA, not used#

hla_run
: Whether to use QUILT to generate posterior state probabilities as part of QUILT-HLA

downsampleToCov
: What coverage to downsample individual sites to. This ensures no floating point errors at sites with really high coverage

minGLValue
: For non-Gibbs full imputation, minimum allowed value in haplotype gl, after normalization. In effect, becomes 1/minGLValue becomes maximum difference allowed between genotype likelihoods

minimum_number_of_sample_reads
: Minimum number of sample reads a sample must have for imputation to proceed. Samples that have fewer reads than this will not be imputed in a given region and all output will be set to missing

nGen
: Number of generations since founding or mixing. Note that the algorithm is relatively robust to this. Use nGen = 4 * Ne / K if unsure

reference_vcf_file
: Path to reference VCF file with haplotypes, matching the reference haplotype and legend file

reference_haplotype_file
: Path to reference haplotype file in IMPUTE format (file with no header and no rownames, one row per SNP, one column per reference haplotype, space separated, values must be 0 or 1)

reference_legend_file
: Path to reference haplotype legend file in IMPUTE format (file with one row per SNP, and a header including position for the physical position in 1 based coordinates, a0 for the reference allele, and a1 for the alternate allele)

reference_sample_file
: Path to reference sample file (file with header, one must be POP, corresponding to populations that can be specified using reference_populations)

reference_populations
: Vector with character populations to include from reference_sample_file e.g. CHB, CHS

reference_phred
: Phred scaled likelihood or an error of reference haplotype. Higher means more confidence in reference haplotype genotypes, lower means less confidence

reference_exclude_samplelist_file
: File with one column of samples to exclude from reference samples e.g. in validation, the samples you are imputing

region_exclude_file
: File with regions to exclude from constructing the reference panel. Particularly useful for QUILT_HLA, where you want to exclude SNPs in the HLA genes themselves, so that reads contribute either to the read mapping or state inference. This file is space separated with a header of Name, Chr, Start and End, with Name being the HLA gene name (e.g. HLA-A), Chr being the chromosome (e.g. chr6), and Start and End are the 1-based starts and ends of the genes (i.e. where we don't want to consider SNPs for the Gibbs sampling state inference)

genetic_map_file
: Path to file with genetic map information, a file with 3 white-space delimited entries giving position (1-based), genetic rate map in cM/Mbp, and genetic map in cM. If no file included, rate is based on physical distance and expected rate (expRate)

nMaxDH
: Integer Maximum number of distinct haplotypes to store in reduced form. Recommended to keep as 2 ** N - 1 where N is an integer greater than 0 i.e. 255, 511, etc

make_fake_vcf_with_sites_list
: Whether to output a list of sites as a minimal VCF, for example to use with GATK 3 to genotype given sites

output_sites_filename
: If make_fake_vcf_with_sites_list is TRUE, optional desired filename where to output sites VCF

expRate
: Expected recombination rate in cM/Mb

maxRate
: Maximum recomb rate cM/Mb

minRate
: Minimum recomb rate cM/Mb

print_extra_timing_information
: Print extra timing information, i.e. how long sub-processes take, to better understand why things take as long as they do

small_ref_panel_block_gibbs_iterations
: What iterations to perform block Gibbs sampling for the Gibbs sampler

`small_ref_panel_gibbs_iterations`
: How many iterations to run the Gibbs sampler for each time it is run (i.e. how many full passes to run the Gibbs sampler over all the reads)

`plot_per_sample_likelihoods`
: Plot per sample likelihoods i.e. the likelihood as the method progresses through the Gibbs sampling iterations

`use_small_eHapsCurrent_tc`
: For testing purposes only

`mspbwtL`
: How many neighouring haplotypes to scan up and down at each grid.

`mspbwtM`
: Minimun long grids matches

`use_mspbwt`
: Use msPBWT to select new haplotypes

`mspbwt_nindices`
: How many mspbwt indices to build

`use_splitreadgl`
: Use split real GL in hap selection and imputation

`override_use_eMatDH_special_symbols`
: Not for general use. If NA will choose version appropriately depending on whether a PBWT flavour is used.

`use_hapMatcherR`
: Used for nMaxDH less than or equal to 255. Use R raw format to hold hapMatcherR. Lowers RAM use

`shard_check_every_pair`
: When using shard gibbs sampler, whether to check every pair of SNPs, or not

`use_eigen`
: Use eigen library for per haploid full li and stephens pass of full haplotype reference panel

`impute_rare_common`
: Whether to use common SNPs first for imputation, followed by a round of rare imputation

`rare_af_threshold`
: Allele frequency yhreshold under which SNPs are considered rare, otherwise they are considered common

`make_heuristic_plot`
: Whether to make a plot for understanding heuristic performance

`heuristic_approach`
: Which heuristic to use

`use_list_of_columns_of_A`
: If when using mspbwt, use columns of A rather than the whole thing, to speed up this version

`calculate_gamma_on_the_fly`
: If when calculating genProbs, calculate gamma on the fly rather than saving #' @return Results in properly formatted version

Author(s)

Robert Davies

QUILT_HLA *QUILT_HLA*

Description

QUILT_HLA

Usage

```
QUILT_HLA(
  bamlist,
  region,
  dict_file,
  hla_gene_region_file = NULL,
  outputdir = "",
  summary_output_file_prefix = "quilt.hla.output",
  nCores = 1,
  prepared_hla_reference_dir = "",
  quilt_hla_haplotype_panelfile = "",
  final_output_RData_file = NA,
  write_summary_text_files = TRUE,
  overrideoutput = FALSE,
  nGibbsSamples = 15,
  n_seek_iterations = 3,
  quilt_seed = NA,
  chr = "chr6",
  quilt_buffer = 500000,
  quilt_bqFilter = 10,
  summary_best_alleles_threshold = 0.99,
  downsampleToCov = 30
)
```

Arguments

bamlist	Path to file with bam file locations. File is one row per entry, path to bam files. Bam index files should exist in same directory as for each bam, suffixed either .bam.bai or .bai
region	HLA region to be analyzed, for example A for HLA-A
dict_file	Path to dictionary file for reference build
hla_gene_region_file	For reference packages built after QUILT 1.0.2, this is not used. For older reference packages, this is needed, and is a path to file with gene boundaries. 4 columns, named Name Chr Start End, with respectively gene name (e.g. HLA-A), chromsome (e.g. chr6), and 1 based start and end positions of gene
outputdir	What output directory to use. Otherwise defaults to current directory

summary_output_file_prefix
: Prefix for output text summary files

nCores
: How many cores to use

prepared_hla_reference_dir
: Output directory containing HLA reference material necessary for QUILT HLA

quilt_hla_haplotype_panelfile
: Prepared HLA haplotype reference panel file

final_output_RData_file
: Final output RData file path, if desired

write_summary_text_files
: Whether to write out final summary text files or not

nGibbsSamples
: How many QUILT Gibbs samples to perform

n_seek_iterations
: How many seek iterations to use in QUILT Gibbs sampling

quilt_seed
: When running QUILT Gibbs sampling, what seed to use, optionally

chr
: What chromosome, probably chr6 or maybe 6

quilt_buffer
: For QUILT Gibbs sampling, what buffer to include around center of gene

quilt_bqFilter
: For QUILT Gibbs sampling, do not consider sequence information if the base quality is below this threshold

summary_best_alleles_threshold
: When reporting results, give results until posterior probability exceeds this value

downsampleToCov
: For imputing states specifically using QUILT, what coverage to downsample individual sites to. This ensures no floating point errors at sites with really high coverage. This is not used in the direct read mapping

Value

Results in properly formatted version

Author(s)

Robert Davies

`QUILT_HLA_prepare_reference` *QUILT_HLA_prepare_reference*

Description

QUILT_HLA_prepare_reference

Usage

```
QUILT_HLA_prepare_reference(
  outputdir,
  nGen,
  hla_types_panel,
  ipd_igmt_alignments_zip_file,
  ref_fasta,
  refseq_table_file,
  full_regionStart,
  full_regionEnd,
  buffer,
  region_exclude_file,
  genetic_map_file = "",
  reference_haplotype_file,
  reference_legend_file,
  reference_sample_file,
  reference_exclude_samplelist_file = "",
  reference_exclude_samples_for_initial_phasing = FALSE,
 all_hla_regions = c("A", "B", "C", "DMA", "DMB", "DOA", "DOB", "DPA1", "DPA2", "DPB1",
  "DPB2", "DQA1", "DQA2", "DQB1", "DRA", "DRB1", "DRB3", "DRB4", "DRB5", "E", "F", "G",
  "HFE", "H", "J", "K", "L", "MICA", "MICB", "N", "P", "S", "TAP1", "TAP2", "T", "U",
   "V", "W", "Y"),
  hla_regions_to_prepare = c("A", "B", "C", "DQA1", "DQB1", "DRB1"),
  chr = "chr6",
  minRate = 0.1,
  nCores = 1
)
```

Arguments

outputdir	What output directory to use
nGen	Number of generations since founding or mixing. Note that the algorithm is relatively robust to this. Use nGen = 4 * Ne / K if unsure, where K is number of haplotypes.
hla_types_panel	Path to file with 1000 Genomes formatted HLA types (see example for format details)
ipd_igmt_alignments_zip_file	Path to zip file with alignments from IPD-IGMT (see README and example for more details)
ref_fasta	Path to reference genome fasta
refseq_table_file	Path to file with UCSC refseq gene information (see README and example for more details)
full_regionStart	When building HLA full reference panel, start of maximal region spanning all HLA genes. The 1-based position x is kept if regionStart <= x <= regionEnd

`full_regionEnd` As above, but end of maximal region spanning all HLA genes

`buffer` Buffer of region to perform imputation over. So imputation is run form regionStart-buffer to regionEnd+buffer, and reported for regionStart to regionEnd, including the bases of regionStart and regionEnd

`region_exclude_file`
File with regions to exclude from constructing the reference panel. Particularly useful for QUILT_HLA, where you want to exclude SNPs in the HLA genes themselves, so that reads contribute either to the read mapping or state inference. This file is space separated with a header of Name, Chr, Start and End, with Name being the HLA gene name (e.g. HLA-A), Chr being the chromosome (e.g. chr6), and Start and End are the 1-based starts and ends of the genes (i.e. where we don't want to consider SNPs for the Gibbs sampling state inference)

`genetic_map_file`
Path to file with genetic map information, a file with 3 white-space delimited entries giving position (1-based), genetic rate map in cM/Mbp, and genetic map in cM. If no file included, rate is based on physical distance and expected rate (expRate)

`reference_haplotype_file`
Path to reference haplotype file in IMPUTE format (file with no header and no rownames, one row per SNP, one column per reference haplotype, space separated, values must be 0 or 1)

`reference_legend_file`
Path to reference haplotype legend file in IMPUTE format (file with one row per SNP, and a header including position for the physical position in 1 based coordinates, a0 for the reference allele, and a1 for the alternate allele)

`reference_sample_file`
Path to reference sample file (file with header, one must be POP, corresponding to populations that can be specified using reference_populations)

`reference_exclude_samplelist_file`
File with one column of samples to exclude from final reference panels. These samples can be removed at one of two points, depending on reference_exclude_samples_for_initial_phasi
If reference_exclude_samples_for_initial_phasing is FALSE, then these samples, if present, are used for the initial phasing and allele assignment. If TRUE, then they are removed immediately

`reference_exclude_samples_for_initial_phasing`
See help for reference_exclude_samplelist_file

`all_hla_regions`
Character vector of all HLA genes for which IPD-IMGT files are available for

`hla_regions_to_prepare`
Character vector of HLA genes to prepare for imputation

`chr` What chromosome to run (probably chr6 or similar)

`minRate` Minimum recomb rate cM/Mb

`nCores` How many cores to use

Value

Results in properly formatted version

Author(s)

Robert Davies

QUILT_prepare_reference

QUILT_prepare_reference

Description

QUILT_prepare_reference

Usage

```
QUILT_prepare_reference(
  outputdir,
  chr,
  nGen,
  regionStart = NA,
  regionEnd = NA,
  buffer = NA,
  output_file = "",
  reference_vcf_file = "",
  reference_haplotype_file = "",
  reference_legend_file = "",
  reference_sample_file = "",
  reference_populations = NA,
  reference_phred = 30,
  reference_exclude_samplelist_file = "",
  region_exclude_file = "",
  genetic_map_file = "",
  nMaxDH = NA,
  tempdir = NA,
  make_fake_vcf_with_sites_list = FALSE,
  output_sites_filename = NA,
  expRate = 1,
  maxRate = 100,
  minRate = 0.1,
  use_mspbwt = FALSE,
  mspbwt_nindices = 4L,
  override_use_eMatDH_special_symbols = NA,
  use_hapMatcherR = TRUE,
  impute_rare_common = FALSE,
  rare_af_threshold = 0.001,
  use_list_of_columns_of_A = TRUE
)
```

Arguments

outputdir	What output directory to use
chr	What chromosome to run. Should match BAM headers
nGen	Number of generations since founding or mixing. Note that the algorithm is relatively robust to this. Use nGen = 4 * Ne / K if unsure
regionStart	When running imputation, where to start from. The 1-based position x is kept if regionStart <= x <= regionEnd
regionEnd	When running imputation, where to stop
buffer	Buffer of region to perform imputation over. So imputation is run form regionStart-buffer to regionEnd+buffer, and reported for regionStart to regionEnd, including the bases of regionStart and regionEnd
output_file	Path to output RData file containing prepared haplotypes (has default value that works with QUILT)
reference_vcf_file	Path to reference haplotype file in VCF format (values must be 0 or 1)
reference_haplotype_file	Path to reference haplotype file in IMPUTE format (file with no header and no rownames, one row per SNP, one column per reference haplotype, space separated, values must be 0 or 1)
reference_legend_file	Path to reference haplotype legend file in IMPUTE format (file with one row per SNP, and a header including position for the physical position in 1 based coordinates, a0 for the reference allele, and a1 for the alternate allele)
reference_sample_file	Path to reference sample file (file with header, one must be POP, corresponding to populations that can be specified using reference_populations)
reference_populations	Vector with character populations to include from reference_sample_file e.g. CHB, CHS
reference_phred	Phred scaled likelihood or an error of reference haplotype. Higher means more confidence in reference haplotype genotypes, lower means less confidence
reference_exclude_samplelist_file	File with one column of samples to exclude from reference samples e.g. in validation, the samples you are imputing
region_exclude_file	File with regions to exclude from constructing the reference panel. Particularly useful for QUILT_HLA, where you want to exclude SNPs in the HLA genes themselves, so that reads contribute either to the read mapping or state inference. This file is space separated with a header of Name, Chr, Start and End, with Name being the HLA gene name (e.g. HLA-A), Chr being the chromosome (e.g. chr6), and Start and End are the 1-based starts and ends of the genes (i.e. where we don't want to consider SNPs for the Gibbs sampling state inference)

genetic_map_file
: Path to file with genetic map information, a file with 3 white-space delimited entries giving position (1-based), genetic rate map in cM/Mbp, and genetic map in cM. If no file included, rate is based on physical distance and expected rate (expRate)

nMaxDH
: Integer Maximum number of distinct haplotypes to store in reduced form. Recommended to keep as 2 ** N - 1 where N is an integer greater than 0 i.e. 255, 511, etc

tempdir
: What directory to use as temporary directory. If set to NA, use default R tempdir. If possible, use ramdisk, like /dev/shm/

make_fake_vcf_with_sites_list
: Whether to output a list of sites as a minimal VCF, for example to use with GATK 3 to genotype given sites

output_sites_filename
: If make_fake_vcf_with_sites_list is TRUE, optional desired filename where to output sites VCF

expRate
: Expected recombination rate in cM/Mb

maxRate
: Maximum recomb rate cM/Mb

minRate
: Minimum recomb rate cM/Mb

use_mspbwt
: Build mspbwt indices to be used in imputation

mspbwt_nindices
: How many mspbwt indices to build

override_use_eMatDH_special_symbols
: Not for general use. If NA will choose version appropriately depending on whether a PBWT flavour is used.

use_hapMatcherR
: Used for nMaxDH less than or equal to 255. Use R raw format to hold hapMatcherR. Lowers RAM use

impute_rare_common
: Whether to use common SNPs first for imputation, followed by a round of rare imputation

rare_af_threshold
: Allele frequency threshold under which SNPs are considered rare, otherwise they are considered common.

use_list_of_columns_of_A
: If when using mspbwt, use columns of A rather than the whole thing, to speed up this version

Value

Results in properly formatted version

Author(s)

Robert Davies

Index