

Package: cpdirection (via r-universe)

October 10, 2026

Type Package

Title Predict the Direction of CpG-Gene Expression Associations

Version 2.99.5

Description Assigns a direction to the association between DNA methylation at a CpG site and the expression of a nearby gene, for CpGs that have no measured expression quantitative trait methylation (eQTM) record. Predictions come from a precomputed table of 3,105,697 CpG-gene pairs covering 761,445 CpGs, assembled from four sources of evidence: directly measured eQTMs, summary-data Mendelian randomisation using genetic instruments, gradient-boosted models of genomic context and local catalogue structure fitted separately in blood, nasal epithelium and solid tissue, and a distance prior. Each call returns a direction, a calibrated probability, an evidence tier and an expected accuracy, and abstains where the evidence is insufficient. Predictions estimate direction conditional on the CpG being an eQTM; they do not establish that an effect exists.

License MIT + file LICENSE

Encoding UTF-8

Depends R (>= 4.5.0)

Imports data.table (>= 1.14.0), utils, stats,
TxDb.Hsapiens.UCSC.hg19.knownGene, org.Hs.eg.db,
GenomicFeatures, GenomicRanges, AnnotationDbi, cpdirectionData

Suggests testthat (>= 3.0.0), gt (>= 0.9.0), knitr, rmarkdown

VignetteBuilder knitr

biocViews DNAMethylation, GeneExpression, GeneRegulation, GeneTarget,
Epigenetics, Annotation

LazyData false

URL <https://github.com/teindor/cpdirection>,
<https://doi.org/10.17605/OSF.IO/U3VFK>

BugReports <https://github.com/teindor/cpdirection/issues>

Config/roxygen2/version 8.1.0
Config/pak/sysreqs
make libbz2-dev libicu-dev liblzma-dev libpng-dev libxml2-dev libssl-dev xz-utils zlib1g-dev
Repository https://biocstaging.r-universe.dev
Date/Publication 2026-09-02 14:15:15 UTC
RemoteUrl https://github.com/BiocStaging/cpgdirection
RemoteRef HEAD
RemoteSha 3e1d93c7a91503cf949dbd3fac28c60f474b2900

Contents

cpg_brain_bridge	2
cpg_direction_all_tissues	4
cpg_direction_universal	5
cpg_expression_direction	6
cpg_gene_pairs	10
cpgd_accuracy_table	13
cpgd_brain_directions	14
cpgd_bridge_deliverable	14
cpgd_build_gene_tss	15
cpgd_build_manifest_genes	16
cpgd_cpg_positions	17
cpgd_direction_balance	18
cpgd_gene_tss	19
cpgd_has_data	19
cpgd_lookup_info	20
cpgd_manifest_genes	21
cpgd_measured_eqtms	22
cpgd_parse_check	23
cpgd_probe_qc	23
cpgd_report	24
cpgd_selftest	25
cpgd_smr_directions	26
Index	28

cpg_brain_bridge	<i>What does a peripheral CpG measurement license about the brain?</i>
------------------	--

Description

For each input CpG: the brain causal direction and target gene(s) where a tier-T1/T2 estimate exists, the peripheral-brain concordance evidence for the tissue you measured, the deliverable grade where the CpG reaches the graded set, and otherwise a genome-wide synthesised saliva-bridge score.

Usage

```
cpg_brain_bridge(cpgs, tissue = c("saliva", "blood", "buccal"))
```

Arguments

cpgs	CpG identifiers in any of the formats <code>cpg_expression_direction()</code> accepts.
tissue	Which peripheral tissue you measured: "saliva" (default), "blood" or "buccal".

Value

A data.table of class `cpgd_bridge`: per CpG, `bridge_grade`, `bridge_usable` (concordance evidence in YOUR tissue), `rho_epic/rho_450k` (IMAGE-CpG, $n = 27$), `amaze_replicated`, `hannon_cortical_rho` (blood, $n = 75$), `sommerer_buccal_R` ($n = 120$), `ensemble_pctile` (genome-wide synthesised saliva score), `strict_qc`, and the brain layer's `t1_genes/t1_directions`. Where a gene was named in the input, `requested_gene`, `requested_gene_direction` and `requested_gene_tier` answer for that gene specifically.

The three caveats this function will keep repeating

1. Directions are BRAIN-derived. Peripheral eQTM directions agree with brain only ~65% of the time and are never used here.
2. Peripheral-brain concordance couplings rest on $n = 27$ living-brain donors; per-CpG estimates have a cross-resource reliability of $r \sim 0.2-0.4$. Grades exist precisely because single estimates cannot be trusted alone.
3. Sex-chromosome and discontinued probes produce replicable artefactual concordance; `strict_qc` is the guard. This function never returns an X/Y probe as bridge-usable.

Co-regulation columns

`co_up` and `co_down` list T1 partner genes this CpG's methylation moves up and down respectively - genes within one column rise and fall together (causal co-direction, from the brain layer itself); a CpG with both columns filled is bivalent and its partners move in opposite directions. `hic_genes` lists genes contacted by the same adult-cortex enhancer the CpG sits in (PsychENCODE); candidate co-regulation by physical contact, with no direction attached - the evidence class that corroborated causal targets 2.5x over nearest genes.

Examples

```
cpg_brain_bridge("cg06846259", tissue = "saliva") # POMC, grade A++
```

cpg_direction_all_tissues

Direction across all tissues, with a cross-tissue consensus

Description

Runs the prediction in every packaged tissue and returns one row per CpG with a column per tissue plus an aggregate. This is usually what you want when the sampled tissue is a proxy for something else: saliva, blood and buccal epithelium are rarely the tissue of interest, and a direction that holds across several tissues is more useful than one tied to whichever proxy happened to be collected.

Usage

```
cpg_direction_all_tissues(
  cpgs,
  genes = NULL,
  min_tissues = 2L,
  allow_nearest_gene = FALSE,
  verbose = TRUE
)
```

Arguments

cpgs	Character vector of CpG identifiers, as for cpg_expression_direction .
genes	Optional target genes, same length as cpgs.
min_tissues	Minimum number of tissues that must return a call before a consensus is reported. Default 2.
allow_nearest_gene	Passed through; see cpg_expression_direction .
verbose	Print a summary. Default TRUE.

Details

The aggregate is justified empirically rather than assumed. Among CpG-gene pairs measured in at least two of the packaged catalogues, direction agrees in 90.5% of cases (1,267 pairs), and excess concordance on shared pairs runs +0.41 to +0.46 within and between blood and epithelium. Where tissues disagree, consensus_direction is NA and tissue_agreement records how badly.

Value

A data.table with, per CpG: a direction, confidence and target gene for each tissue; consensus_direction, consensus_confidence, n_tissues_calling and tissue_agreement; and the measured-evidence columns described in [cpgd_measured_eqtms](#).

Interpretation

consensus_direction is the direction agreed by the calling tissues. It is **not** a claim about any one tissue, and it is not a substitute for measured RNA. It is the best available answer to "if this CpG affects expression of this gene anywhere we have looked, which way does it go".

A CpG can legitimately have opposite directions for different genes, and 20.3% of CpGs with measured evidence are associated with more than one gene. Read the gene columns, not just the direction.

Examples

```
res <- cpg_direction_all_tissues(c("cg00050692", "cg00000029"),
                                verbose = FALSE)
res[, c("cpg_id", "best_direction", "best_evidence"), with = FALSE]
```

cpg_direction_universal

Direction for any CpG-gene pair, from distance alone

Description

Catalogue-based prediction only covers genes some catalogue measured, which is why genes such as CRH, MC2R or HTR2C return NO_EXPRESSION_ANCHOR — they are not transcribed in blood, so no blood catalogue contains them. This function drops that restriction and predicts from distance alone, for any pair whose coordinates you can supply.

Usage

```
cpg_direction_universal(
  cpgs,
  genes,
  gene_tss = NULL,
  cpg_pos = NULL,
  tissue = "all",
  max_dist = 1e+06,
  verbose = TRUE
)
```

Arguments

cpgs	Character vector of CpG identifiers.
genes	Target genes, same length as cpgs. Optional: if omitted, the symbol is taken from the identifier where one is present, as in "cg01961214_TC21_CRHR1".
gene_tss	Optional. Defaults to <code>cpgd_gene_tss</code> , which builds the annotation from TxDb on first use and caches it. Pass a path or a data.frame with gene, chr, tss to override.

cpg_pos	Optional data.frame with cpg_id, chr, pos in hg19. Defaults to the packaged EPIC v2 positions, 930,178 probes.
tissue	Which tissue's curve to use, or "all" for one column per tissue plus a consensus.
max_dist	Pairs further apart than this are not called. Default 1e6.
verbose	Print a summary.

Details

It is deliberately weaker and says so. The distance-only relationship reaches AUC 0.616 within cohort against 0.793 for the full blood model, because it uses none of the neighbourhood or gene-level evidence. Everything it returns carries evidence_tier = "U".

Value

A data.table with the CpG, gene, distance, and a direction and probability per tissue.

What this is and is not

The estimate is the empirical fraction of positive associations at that distance in that tissue class, isotonically smoothed and reported with its 95% interval. It is a prior conditioned on distance, not a statement about your specific CpG, and like everything else here it is conditional on the CpG affecting expression at all. Use it to orient a hypothesis, not to settle one.

Examples

```
if (cpgd_has_data("cpg_positions_hg19") &&
    requireNamespace("TxDb.Hsapiens.UCSC.hg19.knownGene", quietly = TRUE)) {
  cpg_direction_universal(c("cg03405789", "cg08215831"),
                        genes = c("CRH", "CRH"))
}
```

cpg_expression_direction

Predict the direction of a CpG's effect on gene expression

Description

The main entry point. By default it runs every method the package has and returns one row per CpG carrying all of them, plus a single best_direction column holding the answer from the strongest evidence available for that CpG.

Usage

```
cpg_expression_direction(
  cpgs,
  genes = NULL,
  tissue = "all",
  min_tissues = 2L,
  min_confidence = NULL,
  allow_nearest_gene = FALSE,
  universal = TRUE,
  min_distance_info = 0.02,
  target_tissue = NULL,
  html_out = NULL,
  verbose = TRUE
)
```

Arguments

cpgs	Character vector of CpG identifiers, with or without suffixes. A single file path to a text or CSV file is also accepted.
genes	Optional character vector of target genes, the same length as cpgs. Where identifiers already carry the symbol, as in "cg02079741_TC21_POMC", it is parsed automatically.
tissue	"all" (default) runs everything and adds the best_* columns. "blood", "nasal_epithelium" or "solid_tissue" returns that tissue alone.
min_tissues	Tissues that must agree before consensus_direction is reported. Default 2.
min_confidence	Optional numeric in [0, 1] overriding the tier-specific abstention thresholds.
allow_nearest_gene	If a gene is requested but no pairing for it is held, the nearest catalogued gene is a <i>different</i> gene. By default those rows abstain. See cpgd_accuracy_table .
universal	Include the distance-only fallback. Default TRUE. Requires the Bioconductor annotation packages, and builds a TSS table on first use. Set FALSE to skip it.
min_distance_info	Floor on the pooled distance probability, as $ p - 0.5 * 2$, below which a unanimous distance call is withheld. Default 0.02. It is a floor, not the main gate: whether a distance call is made at all is decided by whether the three tissue curves agree in sign.
target_tissue	Optional, one of "blood", "nasal_epithelium" or "solid_tissue". Where the tissues disagree, resolve using this tissue instead of abstaining. Declare it on biological grounds before looking at the votes; the resulting rows are worth that tissue's accuracy, not the consensus figure, and are still flagged in direction_uncertain. Setting it also fills any row that would otherwise abstain with that tissue's sign, labelled targeted_last_resort, so best_direction comes back complete.
html_out	Optional. TRUE writes cpgdirection_results.html into the working directory; a string is used as the file name. See cpgd_report .
verbose	Print a summary. Default TRUE.

Details

Three sources are consulted, in this order:

1. a directly measured eQTM for the requested gene, where one exists;
2. the catalogue models, run in all three packaged tissues, with a consensus where they agree;
3. distance to the transcription start site, which applies to any CpG-gene pair and is much weaker.

Passing a specific tissue instead returns the single-tissue result, with the column set documented under *Value* in earlier versions.

One row per CpG, by design. `cpg_expression_direction()` is the convenience one-result-per-CpG interface: it returns one requested/best target per CpG. For automatic discovery of every supported target/co-effect and pair-specific direction evidence, use `cpg_gene_pairs` – a CpG can have several target genes, with legitimately opposite directions, and this interface by construction shows only one of them. DMSA and any other many-to-many consumer must use the pair API.

Value

With `tissue = "all"`, a `data.table` of class `cpgd_full`, one row per unique CpG:

best_direction +1, -1 or NA: the answer from the strongest evidence available

best_evidence which source it came from — `measured`, `catalogue_consensus`, `catalogue_single`, `distance_only`, `distance_tissue_conflict`, `distance_uninformative`, `tissue_conflict` or `no_evidence`

best_confidence, **best_expected_accuracy** what that answer is worth

consensus_direction, **n_tissues_calling**, **tissue_agreement** the catalogue layer

gene_*, **dir_***, **conf_***, **tier_***, **status_*** per tissue

gene_universal, **dist_universal**, **dir_universal**, **p_universal** the distance layer

smr_direction, **smr_tier**, **smr_p**, **smr_n_instruments**, **smr_gene_dist**, **smr_agreement** the Mendelian randomisation layer

smr_gene the gene the SMR direction refers to, never left to inference

smr_gene_match whether `smr_gene` is the same gene the rest of the row concerns. Only matching rows may supply `best_direction`; non-matching rows are reported in full but describe a different pair. Only 58.6% of CpGs present in both the SMR table and a catalogue share even one gene between them, so this is the common case rather than an edge case

smr_heidi_status, **smr_p_heidi** the HEIDI heterogeneity test: "pass", "fail" or "not_tested" (fewer than three instruments survived LD pruning). Reported, never acted on: on these data HEIDI did not separate concordant from discordant directions, and filtering on it would discard the best-instrumented pairs first. See `?cpgd_smr_directions`

smr_in_table whether the CpG has any SMR evidence at all, for any gene. Distinguishes "no instrument exists" from "the instrument points at a different gene", which are different facts with different remedies

measured_genes, **annotation_mismatch**, **mismatch_note** measured evidence, and whether it concerns a gene other than the one requested

The two alternative codings

Where the evidence leaves a sign nearly free to flip — a distance prior at $p = 0.5009$, say — `best_direction` abstains. That is honest but leaves nothing to analyse, so two further columns bracket the ambiguity: `best_direction_filled` takes the point estimate for those rows and `best_direction_flipped` takes its opposite. `direction_uncertain` flags which rows differ.

These are a sensitivity analysis, not a menu. The question they answer is "does my downstream result survive both codings", not "which coding gives me a result". Picking whichever column fits the data better is selecting on the outcome: it doubles the effective number of tests and biases towards significance. If a finding holds under both, the ambiguity does not matter. If it does not, the finding rests on coin flips and should be reported as unresolved.

Why the distance layer reports conflicts

The three tissue curves are not parallel. Blood asymptotes at 0.449 and never crosses 0.5 at any distance; nasal epithelium reaches 0.594 and solid tissue 0.861. Beyond roughly 20 kb they therefore disagree, often sharply.

Averaging them produces a pooled probability near 0.50 that looks like ignorance but is nothing of the kind: blood is saying -1 and solid tissue $+1$, both with confidence. Those rows are reported as `distance_tissue_conflict`, with `dist_dir_blood`, `dist_dir_nasal` and `dist_dir_solid` exposed so the tissue relevant to your question can be read directly. A distance call enters `best_direction` only when all three curves agree.

A structural consequence worth knowing. The blood curve peaks at 0.449. Because a distance call requires unanimity, and blood never reaches 0.5, `distance_only` can never return $+1$ for any CpG at any distance. Every direction that layer contributes is -1 by construction. This is a property of the fitted curves, not of your data, and it is the reason a panel dominated by distance calls will look uniformly negative. Treat `distance_only` as a one-way instrument and read `dist_dir_solid` where the positive signal would be.

Declaring a target tissue

Blood holds an effective veto over distance calls, because unanimity is required and blood never votes $+1$. That is correct when the question is about blood and wrong when it is not. A saliva-derived panel of brain-expressed genes is asking about brain, and blood's opinion should not silence it.

`target_tissue` resolves conflicts using a named tissue. It is tissue **selection**, not voting: name the tissue first, on the biology, then read what it says. Promoting whichever tissue happens to break the tie the way you hoped is selecting on the outcome, and a majority rule tuned after seeing the votes is the same thing wearing a better hat.

The price is explicit. Rows resolved this way carry that tissue's accuracy: 0.77-0.85 for blood, 0.70-0.87 for nasal epithelium (not externally validated), and 0.55-0.70 for solid tissue, which is tumour, provisional, and whose brain and kidney arms did not exceed their own majority baselines under leave-one-cohort-out. If solid tissue is your brain proxy, that last clause is the one to quote in your limitations.

How to read best_direction

It is the direction supported by the best evidence this package holds for that CpG, not a claim about any one tissue and not a substitute for measured RNA in your own samples. Always read `best_evidence` alongside it: a measured row and a `distance_only` row are not comparable, and the difference in what they are worth is roughly 1.00 against 0.62.

`tissue_conflict` means the tissues returned opposite directions. That is reported as NA rather than resolved, because a weaker method should not break a tie between two stronger ones.

Everything here is conditional on the CpG being an eQTM at all. The catalogues behind the models contain only significant associations, so no output can tell you whether an effect exists. `no_evidence` is a correct answer.

Examples

```
res <- cpg_expression_direction(c("cg02079741_TC21_POMC", "cg000000029"),
                              universal = FALSE, verbose = FALSE)
res[, c("cpg_id", "best_direction", "best_evidence",
        "best_expected_accuracy"), with = FALSE]

# a single tissue returns the per-tissue column set instead
one <- cpg_expression_direction("cg000000029", tissue = "blood",
                              verbose = FALSE)
one[, c("cpg_id", "target_gene", "direction", "evidence_tier", "call"),
     with = FALSE]
```

cpg_gene_pairs	<i>All supported target genes and pair-specific directions for a set of CpGs</i>
----------------	--

Description

The pair-preserving interface. Give it bare CpG identifiers and it first discovers every plausible target gene automatically – from the packaged EPIC v2 manifest annotation, the tissue catalogue, directly measured eQTMs and SMR (plus the brain layer on request, and any gene parsed from the input) – and then resolves direction and evidence separately for every resulting CpG x gene pair. One row per supported pair.

Usage

```
cpg_gene_pairs(
  cpgs,
  genes = NULL,
  tissue = c("blood", "nasal_epithelium", "solid_tissue"),
  gene_mode = c("filter", "pairwise"),
  annotation_mode = c("augment", "strict"),
  include = c("manifest", "lookup", "measured", "smr"),
  include_brain = FALSE,
```

```

probe_qc = c("exclude", "flag", "ignore"),
direction_policy = c("best", "all_evidence"),
universal = TRUE,
sources = NULL,
verbose = TRUE
)

```

Arguments

cpgs	Character vector of CpG identifiers, with or without suffixes ("cg26261055", "cg00645497_BC21_T1_CD38"), a data.frame, or a file path – the same input forms as cpg_expression_direction . Every input row is preserved: nothing is deduplicated by cpg_id alone.
genes	Optional character vector of gene symbols. With the default gene_mode = "filter" it is a set: targets are auto-discovered first and then restricted to genes in this set (any length). With gene_mode = "pairwise" it must be the same length as cpgs and cpgs[i] x genes[i] are evaluated as explicit pairs.
tissue	Tissue context for the catalogue mapping layer and the lookup_* audit columns. Direction resolution consults all three packaged catalogues for consensus, exactly as cpg_expression_direction() does.
gene_mode	"filter" (default) or "pairwise"; see genes. The mode is never inferred from vector lengths – state "pairwise" explicitly when you mean it.
annotation_mode	How a gene parsed from an input suffix is used. "augment" (default): it is one more candidate target alongside the auto-discovered ones. "strict": inputs that carry a suffix gene are restricted to their annotated gene(s); bare inputs still auto-discover.
include	Which mapping sources contribute candidate targets. Any subset of c("manifest", "lookup", "measured", "smr"); default all four.
include_brain	Add brain SMR targets (cpgd_brain_directions) to candidate discovery and report brain_direction/brain_tier per pair. Default FALSE: brain evidence is context-specific and is never mixed into a default peripheral run. Even when included, brain directions are reported alongside and never enter best_direction, which is peripheral; use cpg_brain_bridge for the brain question.
probe_qc	How to treat CpGs whose probes are flagged as unreliable by the packaged QC table (cpgd_probe_qc : cross-hybridizing or degenerate mapping, SNP-contaminated extension base, unverified hg19 position). "exclude" (default) drops CpGs whose <i>every</i> replicate probe carries the general mask – a probe that reads the wrong locus assigns methylation to the wrong gene, and no direction evidence can repair that. "flag" keeps them, marked in probe_masked/probe_mask_reasons. "ignore" skips QC entirely. Exclusions are reported and listed in attr(result, "qc_excluded_cpgs").
direction_policy	"best" (default) returns the pair-level best-evidence record with the standard audit columns. "all_evidence" additionally keeps every per-tissue catalogue and distance column.

universal	Include the distance layer in pair-level resolution. Default TRUE; requires the Bioconductor annotation packages on first use, and degrades gracefully when unavailable.
sources	Advanced/testing: a named list overriding packaged resources – any of manifest, lookup (named list by tissue, or one table for tissue), measured, smr, brain, cpg_pos, gene_tss, distance_curves.
verbose	Print the discovery summary. Default TRUE.

Details

You do not have to supply genes. `genes = NULL` means auto-discover all supported targets; it never means "cannot calculate direction" and it never forces a nearest-gene fallback. Supplying genes is an optional restriction (`gene_mode = "filter"`) or an explicit pair request (`gene_mode = "pairwise"`), never a requirement.

Value

A data.table of class `cpgd_pairs`, one row per CpG x gene pair, with (among others):

input_id, input which submitted row(s) the CpG came from (`input_ids/inputs` collapse several)

cpg_id, target_gene, pair_id the biological key

mapping_sources, mapping_primary, mapping_strength how the gene entered the candidate set: EPICv2_manifest, tissue_lookup, measured_eQTM, SMR, brain_SMR, input_annotation, requested_pair – plus `has_*` flags. Mapping provenance is deliberately distinct from direction evidence: manifest annotation says "Illumina annotates this CpG to this gene", not "this CpG regulates this gene"

best_direction, best_evidence, best_confidence, direction_tier, probability_plus1 the pair's direction record, resolved by the same evidence ladder as `cpg_expression_direction()`, strictly within `cpg_id + target_gene`

usable, abstain_reason pair-specific: one gene of a CpG can be usable while another abstains

n_targets_for_cpg, is_coeffect co-effect accounting

Pair invariance

For a fixed pair the direction record does not depend on how the gene entered the candidate set. `cpg_gene_pairs("cg26261055")` (manifest discovers CRHBP), `cpg_gene_pairs("cg26261055_CRHBP")` (suffix supplies it) and `cpg_gene_pairs("cg26261055", genes = "CRHBP", gene_mode = "pairwise")` return the same `best_direction`, `best_evidence` and confidence for `cg26261055 x CRHBP`; only the mapping provenance differs.

Co-effects

A CpG may legitimately have opposite directions for different target genes; both rows are retained and there is no single CpG-level direction. Abstention is likewise pair-specific.

Examples

```
# hermetic demonstration with injected fixture sources (fast, no data):
fx <- data.frame(cpg_id = "cg10000001", target_gene = "GENEA",
  array = "EPICv2", annotation_source = "Illumina_EPICv2",
  refgene_group = "TSS200")
p <- cpg_gene_pairs("cg10000001", universal = FALSE, verbose = FALSE,
  sources = list(manifest = fx, lookup = NULL,
    measured = NULL, smr = NULL,
    probe_qc = NULL))
p[, c("cpg_id", "target_gene", "mapping_primary", "best_evidence"),
  with = FALSE]

if (cpgd_has_data("lookup_blood_hg19")) {
# real resources: bare CpGs, targets discovered automatically
p <- cpg_gene_pairs(c("cg26261055", "cg06495038", "cg26196496"))

# DMSA-style use: discovered targets restricted to a gene set
p <- cpg_gene_pairs(c("cg26261055", "cg06495038"),
  genes = c("CRHBP", "F2RL1"),
  gene_mode = "filter")
}
```

cpgd_accuracy_table	<i>Expected accuracy by evidence tier</i>
---------------------	---

Description

The validation figures behind the expected_accuracy column, so they can be inspected rather than taken on trust.

Usage

```
cpgd_accuracy_table()
```

Value

A data.frame.

Examples

```
cpgd_accuracy_table()
```

cpgd_brain_directions *Brain causal directions (tier T1/T2)*

Description

The genome-wide brain layer: 220,383 CpG-gene pairs with an SMR causal direction (Brain-mMeta mQTLs as exposure, BrainMeta v2 cis-eQTLs as outcome), graded by a score validated out-of-sample in blood and against the independent Gibbs brain catalogue.

Usage

```
cpgd_brain_directions()
```

Value

A data.table: cpg_id, target_gene, direction, tier, score, n_instruments, instrument_agreement, cpg_gene_dist, hla, inv17q21, p_SMR, p_HEIDI, heidi_status.

What the tiers are worth in brain

Validated against measured brain eQTLs (GSE15745; a cohort verified absent from both SMR arms), tier T1 holds 92-96% across every significance rung of the validation catalogue; tier T2 falls to 67-77%. Accuracies are best supported within ~100 kb (the validating array is promoter-weighted) and are conditional on an association existing. Agreement between independently derived blood and brain directions is only ~65%, so do NOT substitute a peripheral eQTM direction for these.

Examples

```
b <- cpgd_brain_directions()
head(b)
```

cpgd_bridge_deliverable

The graded periphery-to-brain deliverable

Description

728 CpGs where a tier-T1 brain causal direction coexists with robust peripheral-brain methylation concordance ($\rho \geq 0.5$ on BOTH IMAGE-CpG arrays, in at least two peripheral tissues, outside the extended MHC), with per-cohort support flags and grades. Grade A++ (n = 205; 184 saliva-usable) is supported by four independent cohorts.

Usage

```
cpgd_bridge_deliverable()
```

Value

A data.table, one row per CpG; see the paper for column definitions. `strict_qc == FALSE` rows sit on probes failing the stringent probe filter and should be treated as provisional.

Examples

```
d <- cpgd_bridge_deliverable()
head(d)
```

cpgd_build_gene_tss	<i>Build a gene TSS annotation for universal mode</i>
---------------------	---

Description

The catalogue-based prediction only covers genes some catalogue measured. To predict for *any* gene, the package needs transcription start sites, which it does not ship — a gene annotation is large, versioned, and better taken from Bioconductor than frozen into a package.

Usage

```
cpgd_build_gene_tss(
  out = "gene_tss_hg19.csv",
  source = c("txdb", "manifest"),
  manifest_file = NULL
)
```

Arguments

<code>out</code>	Path for the resulting CSV. Default "gene_tss_hg19.csv".
<code>source</code>	"txdb" uses TxDb.Hsapiens.UCSC.hg19.knownGene with org.Hs.eg.db for symbols. "manifest" reads an Illumina manifest instead, via <code>manifest_file</code> .
<code>manifest_file</code>	Path to an Illumina EPIC manifest CSV, used when <code>source = "manifest"</code> .

Details

Run this once. It writes a small table that `cpg_direction_universal` then uses.

TxDb is the more accurate source: real transcription start sites, strand resolved, one canonical position per gene. The Illumina manifest's RefGene fields are convenient but give a gene *region* label per probe rather than a TSS, so distances derived from it are cruder.

Value

Invisibly, the path written. The table has gene, chr, tss and strand.

Examples

```
if (requireNamespace("TxDb.Hsapiens.UCSC.hg19.knownGene", quietly = TRUE) &&
    requireNamespace("org.Hs.eg.db", quietly = TRUE)) {
  cpgd_build_gene_tss(out = tempfile(fileext = ".csv"))
}
```

cpgd_build_manifest_genes

Rebuild the EPIC v2 probe-to-gene annotation from an Illumina manifest

Description

Regenerates the long-form table served by [cpgd_manifest_genes](#) from the official Illumina EPIC v2 manifest CSV (e.g. EPIC-8v2-0_A2.csv). Runtime users never need this; the packaged resource ships with the package. Call it to refresh the table from a newer manifest release or to customise the annotation.

Usage

```
cpgd_build_manifest_genes(
  manifest_file,
  out = "epicv2_probe_gene_annotation.csv.gz"
)
```

Arguments

manifest_file	Path to the Illumina EPIC v2 manifest CSV (may be gzipped).
out	Path for the resulting .csv.gz. Defaults to a file named like the packaged resource, in the working directory. To make a rebuilt table the one cpgd_manifest_genes serves, overwrite the packaged file under inst/extdata/ and reinstall.

Details

The builder:

1. reads the manifest, skipping the [Assay] header preamble and stopping before the [Controls] section;
2. locates the probe identifier column (ILmnID/Name) and the UCSC_RefGene_Name / UCSC_RefGene_Group columns (plus GencodeV41_Name / GencodeV41_Group where present);
3. splits every semicolon-delimited multi-gene field, pairing gene and group tokens per transcript – a mismatched pair count yields NA groups rather than guessed ones, and the field is never reduced to its first gene;
4. uppercases and trims gene symbols, dropping Ensembl-ID-only entries;
5. collapses to one row per CpG x gene, pooling transcript groups and retaining the raw manifest fields for provenance;
6. writes the compressed table.

Value

Invisibly, the path written.

Examples

```
f <- tempfile(fileext = ".csv")
writeLines(c("[Heading]", "Descriptor File Name,EPIC-8v2-0_TEST",
  "[Assay]",
  "IlmnID,Name,CHR,MAPINFO,Strand_FR,UCSC_RefGene_Name,UCSC_RefGene_Group",
  "cg10000001_TC21,cg10000001,chr1,1000,F,GENEA;GENEB,TSS200;TSS1500",
  "[Controls]", "ctl1,negative,0,0"), f)
out <- tempfile(fileext = ".csv.gz")
cpgd_build_manifest_genes(f, out = out)
read.csv(out)
```

cpgd_cpg_positions	<i>Packaged CpG positions (EPIC v2, hg19)</i>
--------------------	---

Description

930,178 probe positions, lifted to hg19 and verified at 99.8 percent exact agreement against the HELIX catalogue. Used by [cpg_direction_universal](#); exported because it is generally useful.

Usage

```
cpgd_cpg_positions()
```

Value

A data.table with cpg_id, chr, pos.

Examples

```
p <- cpgd_cpg_positions()
p[p$cpg_id == "cg00000029", ]
```

cpgd_direction_balance

Is the balance of +1 and -1 calls what it should be?

Description

A sanity check for a result that looks suspiciously one-sided. Inverse associations are the majority everywhere, so a run returning mostly -1 is expected; a run returning *only* -1 may still be correct, and this says which.

Usage

```
cpgd_direction_balance(x = NULL, tissue = "blood")
```

Arguments

x	Optional result from cpg_expression_direction .
tissue	Which tissue's base rate to test against. Default "blood".

Details

Called with no argument it reports the balance in the shipped tables. Called on a result it compares that result's catalogue-derived calls against the relevant base rate with an exact binomial test.

Value

Invisibly, a list of the figures printed.

Why distance-only calls are excluded

The pooled distance curve falls below 0.5 out to roughly 19 kb and rises above it only beyond roughly 300 kb, with an uninformative band between. A panel whose CpGs all sit within a few tens of kilobases of their target therefore **cannot** receive a +1 from the distance layer, no matter how many CpGs it contains. That is a property of the panel's geometry, not evidence about its CpGs, so those calls are not tested here.

Examples

```
cpgd_direction_balance()
```

cpgd_gene_tss	<i>Gene TSS annotation, built once and cached</i>
---------------	---

Description

Derives transcription start sites from TxDb.Hsapiens.UCSC.hg19.knownGene, mapping Entrez identifiers to symbols through org.Hs.eg.db. The result is cached under `tools::R_user_dir("cpgdirection", "cache")`, so the cost is paid once per machine rather than once per session.

Usage

```
cpgd_gene_tss(refresh = FALSE)
```

Arguments

refresh	Rebuild even if a cached copy exists.
---------	---------------------------------------

Details

Called automatically by [cpg_direction_universal](#). Call it directly only to force a rebuild or to inspect the table.

Value

A data.table with gene, chr, tss, strand.

Examples

```
if (requireNamespace("TxDb.Hsapiens.UCSC.hg19.knownGene", quietly = TRUE)) {
  g <- cpgd_gene_tss(); g[gene == "CRH"]
}
```

cpgd_has_data	<i>Is a data layer available in this session?</i>
---------------	---

Description

Reports whether a named data layer can be retrieved right now – from the directory named by `options(cpgdirection.data_dir=)`, from the copy bundled with the package, or from the **cpgdirectionData** ExperimentHub package. Examples, vignettes and analysis scripts use it to skip work that has no data behind it instead of failing.

Usage

```
cpgd_has_data(name = NULL)
```

Arguments

name Layer name, for example "lookup_blood_hg19", "smr_directions" or "epicv2_probe_gene_annotation". The default NULL tests the blood catalogue lookup, the layer most functions reach for first.

Value

A single TRUE or FALSE.

Examples

```
cpgd_has_data()
cpgd_has_data("smr_directions")
```

cpgd_lookup_info	<i>Information about the bundled lookup table</i>
------------------	---

Description

Information about the bundled lookup table

Usage

```
cpgd_lookup_info(tissue = c("blood", "nasal_epithelium", "solid_tissue"))
```

Arguments

tissue Which tissue's table to describe: "blood", "nasal_epithelium" or "solid_tissue".

Value

A named list: tissue, number of pairs, CpGs and genes, genome build, file path and package version.

Examples

```
cpgd_lookup_info("blood")
```

cpgd_manifest_genes	<i>EPIC v2 probe-to-gene annotation (long form, multi-source)</i>
---------------------	---

Description

The packaged long-form EPIC v2 probe-to-gene table: one row per CpG x gene, with every annotated gene preserved and, from 2.5.0, every assignment cross-checked across independent annotation tracks. A mis-assigned CpG sends every downstream direction to the wrong gene, so no single source is trusted alone; the shipped table is the UNION of five tracks with per-track support flags and an agreement count:

- `Illumina_UCSC_RefGene` – the official Illumina EPIC-8v2-0 manifest RefGene annotation;
- `Illumina_GencodeV41` – the manifest’s GENCODE v41 track (symbol-named entries);
- `Zhou_GENCODEv41` – the Zhou lab (SeSAmE) transcript-aware GENCODE v41 annotation, contributing `zhou_dist_tss` (signed bp to the nearest annotated TSS of that gene) and transcript types;
- `Exeter_GENCODEv47` – the Exeter re-annotation (Mallabar-Rimmer et al. 2025, GENCODE release 47): gene-body and TSS features per transcript;
- `Exeter_regulatory` – the Exeter distance-based regulatory assignments (`Promoter_2000bp` / `Enhancer_5000bp` element genes), plus the probe-level `in_genehancer` flag.

Semicolon-delimited multi-gene fields are always fully expanded – never reduced to the first gene – and Ensembl-ID-only entries are excluded (nothing else in the package keys on Ensembl IDs).

Usage

```
cpgd_manifest_genes()
```

Details

`n_annotation_sources` is the number of independent tracks supporting the pair (84.4% of shipped pairs have 2 or more; 54.8% have 3 or more). Downstream consumers such as DMSA can weight or filter on it.

Value

A `data.table` with columns `cpg_id`, `target_gene`, `array`, `annotation_source` (semicolon-joined supporting tracks), `n_annotation_sources`, the per-track logical flags `src_Illumina_UCSC_RefGene`, `src_Illumina_GencodeV41`, `src_Zhou_GENCODEv41`, `src_Exeter_GENCODEv47`, `src_Exeter_regulatory`, plus `refgene_group` (Illumina), `exeter_feature`, `in_genehancer`, `zhou_dist_tss`, `zhou_tx_types`, `manifest_probe_id`, `chr`, `position` and `strand`.

What this table is and is not

Annotation is positional: "this source annotates this CpG to this gene". It does not establish that the CpG regulates the gene, and it is deliberately kept distinct from measured eQTM and SMR evidence, which carry different biological meaning. In [cpg_gene_pairs](#) it contributes candidate targets; the direction evidence for each pair is resolved separately. Probe RELIABILITY (does the probe read the right locus at all?) is likewise a separate question, answered by [cpgd_probe_qc](#).

Coordinates

chr, position and strand are hg38 manifest coordinates, carried for provenance. The package's distance computations use the hg19 positions from [cpgd_cpg_positions](#), not these.

Examples

```
m <- cpgd_manifest_genes()
m[m$cpg_id == "cg26261055", ] # CRHBP, 4 independent sources, TSS -899 bp
```

cpgd_measured_eqtms	<i>Measured CpG-gene associations across the packaged catalogues</i>
---------------------	--

Description

The 25,056 CpG-gene pairs that were directly measured in at least one catalogue, rather than predicted: 18,276 CpGs and 4,997 genes.

Usage

```
cpgd_measured_eqtms()
```

Details

These matter for a specific reason. Nearest-gene annotation is wrong often enough to be a problem: 23.1% of measured pairs involve a gene that is *not* the nearest measured gene to that CpG, and 20.3% of these CpGs have measured associations with more than one gene. A CpG annotated to gene X on the array may have its measured relationship with gene Y.

Value

A data.table with cpg_id, target_gene, tissue, direction and tss_dist.

Examples

```
m <- cpgd_measured_eqtms()
m[m$cpg_id == "cg00000165", ]
```

cpgd_parse_check	Show what the identifier parser extracted
------------------	---

Description

Diagnostic. Given the identifiers you intend to pass to [cpg_expression_direction](#), shows the CpG and gene the parser pulls out of each, so a naming convention that does not parse can be seen rather than guessed at.

Usage

```
cpgd_parse_check(cpgs)
```

Arguments

cpgs Character vector of identifiers, as you would pass them.

Value

A data.table with input, cpg_id, given_gene and given_gene2.

Examples

```
cpgd_parse_check(c("Zcg00335286_TC21_MC2R", "cg00000029", "not_a_probe"))
```

cpgd_probe_qc	EPIC v2 probe reliability (QC) table
---------------	--------------------------------------

Description

Per-CpG probe quality assembled from independent sources: the Zhou lab (SeSAMe) EPIC v2 general mask – degenerate or non-unique mapping, SNP contamination at the extension base, and other design faults – the Garvan (Peters et al. 2024) experimental evidence – WGBS-confirmed cross-hybridization, BLAT-predicted off-targets, mapping mismatches – and a verification of the packaged hg19 position against the Zhou hg19 manifest. The two cross-hybridization angles corroborate each other (64% of WGBS-confirmed off-target probes are also sequence-masked by Zhou) and each catches probes the other misses.

Usage

```
cpgd_probe_qc()
```

Details

A probe that reads the wrong locus assigns methylation to the wrong gene entirely – the one failure mode no downstream direction evidence can repair. `cpg_gene_pairs` therefore consults this table and, by default (`probe_qc = "exclude"`), drops CpGs whose every replicate probe carries the general mask.

Exclusion-grade flags follow the EVERY-replicate rule: `masked_general` (Zhou) and `cross_hybridizing` (Garvan, WGBS-evidenced) are TRUE only when every replicate probe of the CpG is flagged; a CpG with at least one clean replicate is kept (`masked_partial`). `qc_exclude` is their union and is the criterion `probe_qc = "exclude"` acts on. Informational flags use the ANY-replicate rule: `ch_blatt_predicted` (computationally predicted off-targets) and `mapping_flagged` (alignment mismatches) mark rows without excluding them.

Value

A `data.table` with `cpg_id`, `n_probes`, `n_masked_general`, `mask_reasons`, `masked_general`, `masked_partial`, `pos_hg19_verified` (NA where the CpG is absent from the verification source), `cross_hybridizing`, `ch_blatt_predicted`, `mapping_flagged`, `n_offtargets_max`, `qc_exclude` and `qc_sources`.

Examples

```
qc <- cpgd_probe_qc()
head(qc[qc$masked_general == TRUE, ], 3)
```

cpgd_report

Write a results table to HTML

Description

Renders a `cpg_expression_direction()` result as a formatted HTML table and writes it to disk. Uses **gt** when it is installed, and falls back to a plain styled table when it is not, so a long run never fails at the last step for want of an optional package.

Usage

```
cpgd_report(
  x,
  file = "cpgdirection_results.html",
  dir = getwd(),
  title = "CpG to expression direction",
  subtitle = NULL,
  compact = TRUE,
  open = FALSE
)
```

Arguments

x	A result from <code>cpg_expression_direction</code> .
file	File name. A relative name is written inside dir.
dir	Directory to write into. Defaults to the working directory, so <code>setwd()</code> controls where results land.
title	Table title.
subtitle	Table subtitle. NULL generates one from the data.
compact	Show the interpretive columns only. FALSE writes every column, which is wide but complete. Default TRUE.
open	Open the file in a browser when done. Default FALSE.

Value

Invisibly, the path written.

Examples

```
res <- cpg_expression_direction(c("cg02079741_TC21_POMC", "cg00000029"),
                               universal = FALSE, verbose = FALSE)
cpgd_report(res, "hpa_panel.html", dir = tempdir())
```

cpgd_selftest

Verify the installation against a shipped reference

Description

Scores 40 CpGs whose expected output was computed when the package was built and compares the results field by field. Run this once after installing; if it fails, do not use the results.

Usage

```
cpgd_selftest(regenerate = NULL)
```

Arguments

regenerate	Optional path. Writes a fresh reference file from the installed package instead of testing against the shipped one. Use after a deliberate change of behaviour, never to make a failure go away.
------------	--

Details

The check is on `probability_plus1`, the model's actual output, together with the assigned gene and evidence tier. Direction is reported but does not gate the result, because it is the sign of the probability and would only restate it.

Value

Invisibly, TRUE on success. Prints a report either way.

Examples

```
cpgd_selftest()
```

cpgd_smr_directions	<i>Causal CpG-gene directions from two-sample SMR</i>
---------------------	---

Description

413,982 CpG-gene pairs over 103,285 CpGs and 14,056 genes, derived by summary-data Mendelian randomisation with DNA methylation as the exposure (GoDMC blood mQTL meta-analysis, ~32,000 individuals) and gene expression as the outcome (eQTLGen blood cis-eQTL, 31,684 individuals).

Usage

```
cpgd_smr_directions()
```

Details

This layer answers a question the catalogues cannot: it does not ask whether anyone measured this CpG against this gene, but whether a genetic instrument for methylation at the CpG moves expression of the gene. Coverage is therefore not limited to pairs some study happened to test, and the median CpG-gene distance is 135 kb, well beyond the range where the catalogue models have support.

Value

A data.table with cp_g_id, target_gene, direction, smr_tier, p_SMR, n_instruments, instrument_agreement, cp_g_gene_dist, top_instrument, and from 2.2.5 p_HEIDI, nsnp_HEIDI and heidi_status ("pass", "fail", "not_tested").

What the tiers are worth

Validated against the directly measured eQTMs shipped with this package, on the pairs where both exist:

tier	pairs	validated on	concordance
S1	29,456	2,141	95.9% (+/- 0.8)
S2	366,838	6,008	84.9% (+/- 0.9)
S3	17,688	456	70.4% (+/- 4.2)

against a majority-class baseline of 59.8% on the same overlap. S1 requires more than one instrument, all agreeing in sign, at p_SMR < 5e-8. S2 is a single instrument at the same threshold. S3 has

multiple instruments that *disagree*, and scores below S2 accordingly: instrument disagreement is a warning, not extra evidence.

What SMR does and does not establish

A significant SMR result is consistent with methylation affecting expression, and equally with **linkage** – two distinct causal variants in LD, one acting on each trait. Agreement among multiple independent instruments (tier S1) argues against linkage but does not exclude it.

Both datasets are whole blood, so the causal claim is about blood whatever tissue the gene matters in. Palindromic SNPs and unresolvable allele mismatches were dropped rather than assumed; 21,698 instrument-gene comparisons were refused on those grounds.

HEIDI, and why it is reported but not applied

Distinguishing mediation from linkage is what HEIDI is for. It was run on all 413,982 pairs with SMR v1.03 against the 1000 Genomes European reference: 22,421 pairs passed (5.4%), 323,464 failed (78.1%), and 68,097 (16.4%) had fewer than three instruments after LD pruning and could not be tested. **Not tested is not passed** and the two are never merged.

The tiering does **not** condition on HEIDI, because on these data it does not discriminate in the direction that would justify a filter. Against the measured eQTMs, passing pairs were concordant 80.8% of the time and failing pairs 87.3% ($p = 0.002$) – the wrong way round. The cause is power: HEIDI's ability to reject rises with instrument count, and so does the precision of the Wald ratio. Concordance by instruments surviving pruning was 69.9% (<3), 74.8% (3-5), 74.0% (6-10) and 88.0% (11-20), while mean instrument count was 17.3 among failing pairs against 12.0 among passing ones. Filtering on HEIDI would discard the best-instrumented and most accurate pairs first.

The columns are shipped so you can condition on them if you wish. Do so with the above in view.

Examples

```
s <- cpgd_smr_directions()
s[s$cpg_id == "cg18998365", ]
```

Index

cpb_brain_bridge, [2](#), [11](#)
cpb_direction_all_tissues, [4](#)
cpb_direction_universal, [5](#), [15](#), [17](#), [19](#)
cpb_expression_direction, [4](#), [6](#), [11](#), [18](#), [23](#),
[25](#)
cpb_gene_pairs, [8](#), [10](#), [22](#), [24](#)
cpb_accuracy_table, [7](#), [13](#)
cpb_brain_directions, [11](#), [14](#)
cpb_bridge_deliverable, [14](#)
cpb_build_gene_tss, [15](#)
cpb_build_manifest_genes, [16](#)
cpb_cpb_positions, [17](#), [22](#)
cpb_direction_balance, [18](#)
cpb_gene_tss, [5](#), [19](#)
cpb_has_data, [19](#)
cpb_lookup_info, [20](#)
cpb_manifest_genes, [16](#), [21](#)
cpb_measured_eqtms, [4](#), [22](#)
cpb_parse_check, [23](#)
cpb_probe_qc, [11](#), [22](#), [23](#)
cpb_report, [7](#), [24](#)
cpb_selftest, [25](#)
cpb_smr_directions, [26](#)