Genome Biology

RESEARCH

Open Access

The distinct roles of genome, methylation, transcription, and translation on protein expression in *Arabidopsis thaliana* resolve the Central Dogma's information flow

Ziming Zhong[1†], Mark Bailey[2†], Yong-In Kim[3†], Nazanin P. Afsharyan[2,4†], Briony Parker[2], Louise Arathoon[1], Xiaowei Li[2], Chelsea A. Rundle[2], Andrew Behrens[5], Danny Nedialkova[5,6], Gancho Slavov[7], Keywan Hassani-Pak[2], Kathryn S. Lilley[3†], Frederica L. Theodoulou[2†] and Richard Mott[1*†]

[†]Ziming Zhong, Mark Bailey, Yong-In Kim and Nazanin P. Afsharyan contributed equally to this work.

[†]Kathryn S. Lilley, Frederica L. Theodoulou and Richard Mott are joint senior authors.

*Correspondence: r.mott@ucl.ac.uk

[1] Genetics Institute, University College London, London WC1E 6BT, UK
Full list of author information is available at the end of the article

Abstract

Background: We investigate the flow of genetic information from DNA to RNA to protein as described by the Central Dogma in molecular biology, to determine the impact of intermediate genomic levels on plant protein expression.

Results: We perform genomic profiling of rosette leaves in two *Arabidopsis* accessions, Col-0 and Can-0, and assemble their genomes using long reads and chromatin interaction data. We measure gene and protein expression in biological replicates grown in a controlled environment, also measuring CpG methylation, ribosome-associated transcript levels, and tRNA abundance. Each omic level is highly reproducible between biological replicates and between accessions despite their ~1% sequence divergence; the single best predictor of any level in one accession is the corresponding level in the other. Within each accession, gene codon frequencies accurately model both mRNA and protein expression. The effects of a codon on mRNA and protein expression are highly correlated but independent of genome-wide codon frequencies or tRNA levels which instead match genome-wide amino acid frequencies. Ribosome-associated transcripts closely track mRNA levels.

Conclusions: DNA codon frequencies and mRNA expression levels are the main predictors of protein abundance. In the absence of environmental perturbation neither gene-body methylation, tRNA abundance nor ribosome-associated transcript levels add appreciable information. The impact of constitutive gene-body methylation is mostly explained by gene codon composition. tRNA abundance tracks overall amino acid demand. However, genetic differences between accessions associate with differential gene-body methylation by inflating differential expression variation. Our data show that the dogma holds only if both sequence and abundance information in mRNA are considered.

Keywords: Gene-body methylation, Mim-tRNAseq, RNAseq, Ribosome-associated expression, Gene expression, Protein expression, Data-independent acquisition, Genome assembly, Chromatin interaction, Long reads, Central Dogma

Background

Numerous studies in plants, fungi, and animals have found only moderately strong relationships between protein and mRNA expression levels, with correlations typically around 0.5–0.6 [1–3]. This phenomenon is thought to be a consequence of several factors, principally, different rates of synthesis and degradation of mRNA and proteins [4], compounded with buffering and cross-talk between different spatial and temporal contexts of expression, protein length [5], measurement bias and inaccuracy [6], and, potentially, how the data are processed.

Recent advances in genomics technologies have made it possible to assemble genomes almost perfectly, to quantify DNA methylation and other epigenetic marks, and to measure protein and transcript expression accurately at scale. We can now leverage these advances to massively extend the accuracy and depth of "omic" data sets, to dissect their relationships and shed light on the mRNA-protein correlation problem. Specifically, by integrating these data across omic levels we can now test if the flow of information supports the direction of the Central Dogma from lower to higher levels, namely genome → epigenome → transcriptome → proteome (here we have added epigenetics to the Central Dogma's flow between genome and transcriptome; the genomic impacts of the environment are assumed to act via the epigenome). Modeling between omic levels for the same gene addresses inter-level correlations, while modeling across genes within each level reveals factors acting differentially between genes.

If we use lower omic levels (primary DNA sequence features and epigenetic marks) to predict higher levels (transcriptome, proteome), and ultimately phenotype—and thereby follow the Central Dogma—[7], then three questions are particularly relevant.

First, which features of the underlying DNA sequence are most predictive of higher omic levels? Second, how much of the information about protein expression levels encoded in the basal genome sequence is mediated through intermediate epigenetic and transcriptomic levels, and does it pass through multiple causal pathways? Third, in plants, what is the role of constitutive gene-body CpG methylation (gbM) in controlling gene and protein expression? In contrast to environmentally modulated gbM, it has been suggested that constitutive gbM might not impact gene expression in plants at all, although it appears to be evolutionarily conserved [8], under selection [9], and to play a role in adaptation [10]. Furthermore, is it unclear how constitutive gbM in a specific gene is set [11, 12].

Large-scale population-based studies can reveal how genetic and environmental variation impact expression of different omic levels, but if our focus is on modeling the Central Dogma there is a strong case for examining smaller systems where genotype and environment are tightly controlled and in which each omic level is measured in many biological replicates as reproducibly as possible. Here, we employed the second approach.

We first asked if lower omic levels predict higher levels, and whether the information they encode is unique to that level, using the fraction of variation across genes in a focal

omic level that is explained by variation in lower levels as our metric. We then asked if the influence of gbM on higher omic levels is subsumed by the information encoded by genome sequence, specifically in gene codon frequencies, if these codon frequencies affect mRNA and protein expression in similar ways, whether the impact of each type of codon is related to its genome-wide frequency, and how the levels of tRNAs relate to mRNA and protein expression. Finally, we analyzed genetically encoded differences in methylation and expression, to test whether the former are related to the latter, which types of differences are most important, and what this tells us about the causal effects of methylation. Our analysis uncovers some unexpected yet important relationships between omic levels, while showing others are insignificant under the experimental conditions employed here.

Results

We performed detailed genomic profiling of two *Arabidopsis thaliana* accessions: Col-0 (the reference) and Can-0. The latter accession originates from the Canary Islands and is phenotypically adapted to an environment quite different from the central European origin of Col-0 [13]. Can-0 has about double the number of sequence differences from Col-0 compared to most other accessions [14] and has been variously characterized as a relict by [15] and as "admixed" and distinct from four main genetic clusters (Europe, Madeira, Asia, Africa) identified by long-read sequencing of 70 accessions in [16]. Can-0 and Col-0 therefore represent genetically distinct lineages of *Arabidopsis*, so identifying evolutionarily shared characteristics between these accessions, in contrast to those that differ, may answer some of the questions described above.

We re-assembled the accessions' genomes and measured constitutive CpG gbM for each gene. We re-annotated each assembled genome to produce accurate data across omic levels, aiming to eliminate reference bias. To try to eliminate environmental perturbations, we grew multiple biological replicates of each accession under the same climate-controlled, long-day environment in growth chambers. We quantified mRNA, tRNA, ribosome-associated transcripts, and protein abundances in rosette leaves of defined developmental age.

Col-0 and Can-0 genome assembly and annotation

We produced high-quality de novo assemblies of the Col-0 and Can-0 genomes, from a combination of long (HiFi, ONT) and short (Illumina) reads, using Omni-C chromatin interaction data to confirm our assemblies and orient scaffolds (Fig. 1A). Seventeen and five gaps remain in our assemblies of Col-0 and Can-0, respectively, all within rDNA arrays and centromeres (Fig. 1B). We observed excessively high depth of coverage of both ONT and HiFi reads in these repeat regions, suggesting the numbers of tandem repeats in centromeres may be underestimated and, potentially, variable between nuclei (Fig. 1A). Outside of large tandem repeats, Omni-C data indicate that the assemblies are structurally accurate (as shown by the Pretext contact maps in Additional file 1: Fig. S1 (Col-0) and Additional file 1: Fig. S2 (Can-0), as do the BUSCO and QV gene content statistics (in Additional file 2: Table S1)).

We aligned our Col-0 sequence to five published Col-0 assemblies, namely the old reference TAIR10 [13], Col-CEN [17], Col-XJTU [18], Col-Lian [16], and the new community

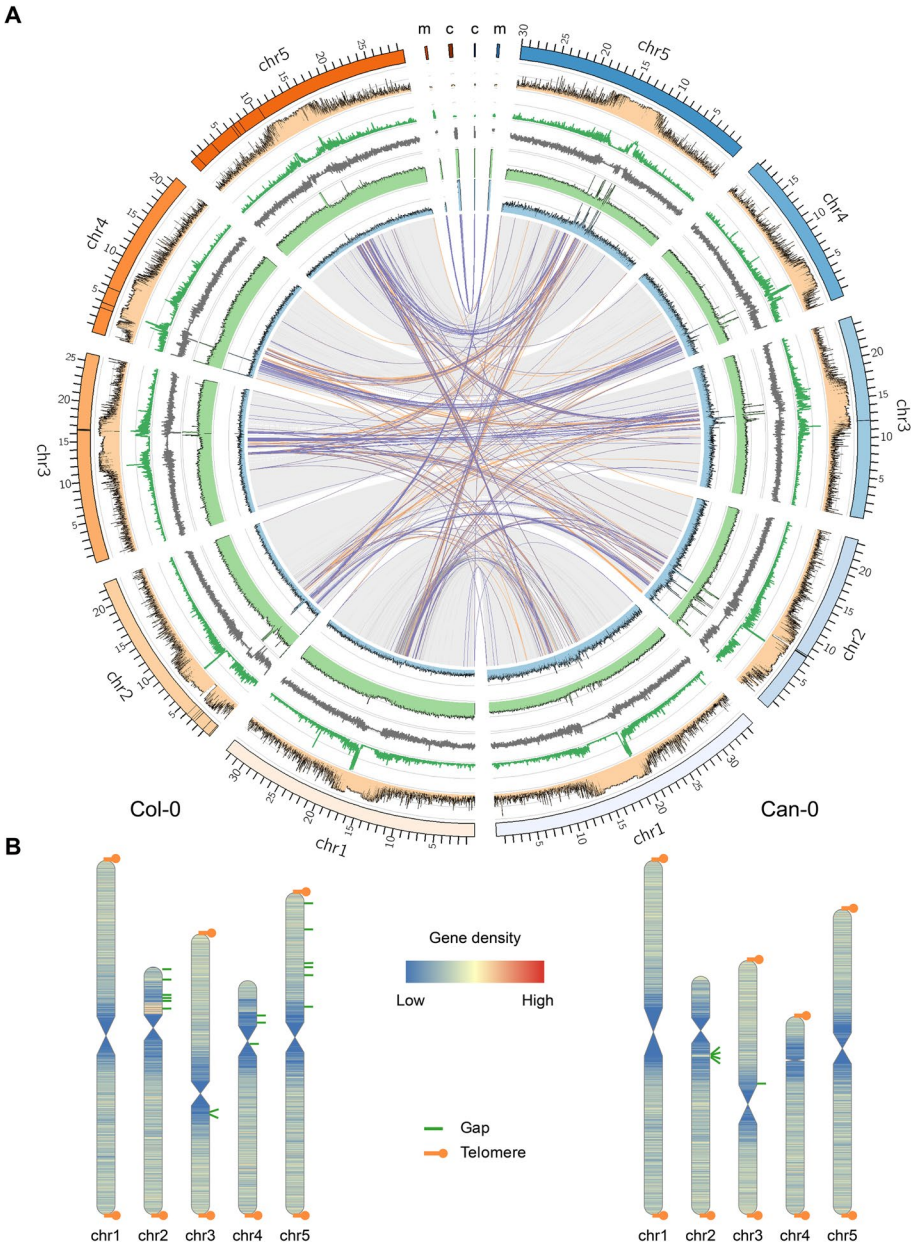

Fig. 1 Assembly of Col-0 and Can-0 genomes. **A** Circos plot comparing the Col-0 (orange set, left) and Can-0 (blue set, right) genomes. Links in the middle show genomic rearrangements. Purple links: inversions, orange links: translocations between different chromosomes. Light blue: HiFi coverage, light green: ONT coverage. Gray line panel: GC content of the genome. Dark green line: repeat density, orange filled panel: percent CpG methylation. **B** Cartoons of chromosome (chr) assemblies of Col-0 (left) and Can-0 (right) showing gene densities, the positions of assembly gaps (green) and indicating where the assemblies reached into the telomeres (orange dots)

reference Col-CC (GenBank reference GCA_028009825.2). The numbers of differences, as computed by dna_diff [19], between our Col-0 assembly and the others are shown in Fig. 2 and are generally small, e.g., there are only about 10,000 SNP differences between the Col-0 assemblies (compared to 682,351 SNP differences between Col-0 and Can-0), and almost all the observed differences are in the numbers of tandem repeats. Some differences

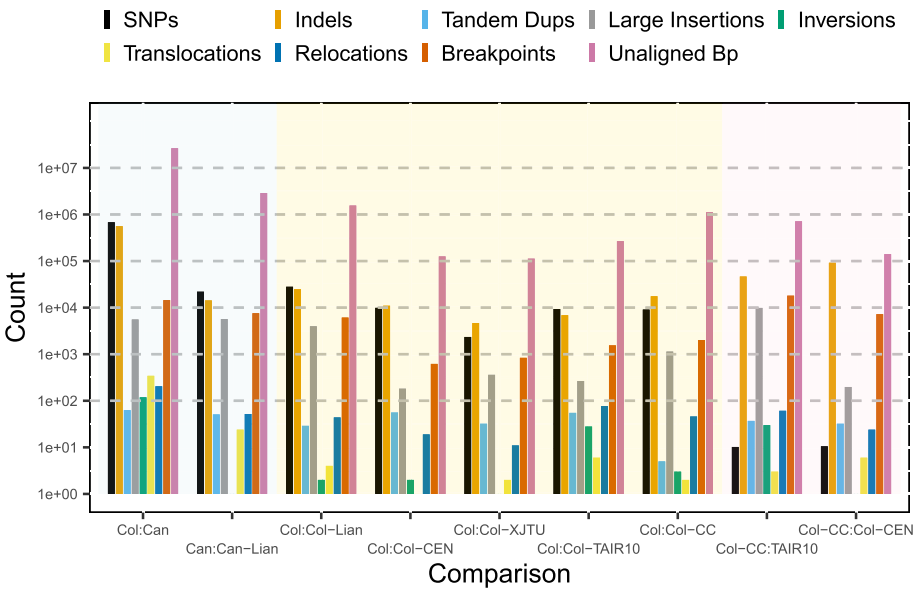

Fig. 2 Counts of differences between our Can-0 (blue background) and Col-0 (yellow background) assemblies and with four other Col-0 and one other Can-0 assemblies, namely Can-Lian, Col-Lian, Col-CEN, Col-XJTU, Col-TAIR, the TAIR10 reference, and Col-CC: the community consensus assembly. Pink background shows comparisons between selected other Col-0 assemblies. Y-axis shows the $\log_{10}$ counts of the respective differences

are likely to be artifacts from different assembly algorithms; indeed, our Col-0 most closely resembles the Col-XJTU assembly, where the same software and similar pipelines were employed (Additional file 2: Table S2). Comparison of Col-CC and Col-Cen shows similar numbers of differences (Fig. 2), so our Col-0 assembly is not unusual.

We also compared our Can-0 assembly to that reported in [16] (named Can-Lian here) and again found similar numbers and patterns of differences as observed between different long-read Col-0 assemblies (Fig. 2, Additional file 2: Table S2). We conclude that all these assemblies have similar accuracies and that the differences represent in part algorithmic artifact. However, we cannot exclude the possibility of small numbers of genuine sequence differences (e.g., under ten thousand SNPs) in the germplasm sequenced, particularly in unstable tandem repeat regions.

Henceforth, "Col-0" and "Can-0" refer to our assemblies and annotations of these accessions, unless otherwise stated. Where we calculate the same statistic in Col-0 and Can-0, the numbers are reported as an ordered pair. For example, our assemblies' lengths (Col-0: 133.23 Mb, Can-0: 133.09 Mb) and N50 values (18.4 Mb, 12.5 Mb) are very similar to those obtained in [17, 18]. We conservatively estimate the sequence divergence of Col-0 and Can-0 at over 1%, based on the total of 1,243,716 SNP and single base indels within their alignable regions; it is greater, if harder to quantify, should the unaligned regions be included.

We annotated both genomes, using ab initio gene prediction applied to both short-read (Illumina) and long-read (Iso-Seq) RNAseq data from rosette leaves sampled at the 9-leaf stage. We annotated 28,763 and 28,532 protein-coding genes. These counts include duplicated genes within each accession. If we exclude duplicates, then there are 24,325 pairs of orthologous genes between Col-0 and Can-0 of which 23,081 are also

annotated in Araport11 [20]. Several hundred genes are unique to each annotation, as shown in the Venn diagram in Fig. 3. The gene and transposable element annotations of the Col-0 and Can-0 genomes along with the predicted mRNA and peptide sequences are available from Figshare [21].

We annotated 47,596 and 45,756 alternatively spliced isoforms (Additional file 3: Table S3). Taking the primary isoform in each accession, in total 20.3% of Col-0:Can-0 orthologous coding sequences are identical at the amino acid level, and 52.7% differ at no more than six codons (see below). However, 7.0% differ by more than 100 codons. Since 35.1% of orthologous gene pairs have different numbers of annotated isoforms, there is some ambiguity in these statistics.

mRNA and protein expression levels are highly reproducible but relatively poorly correlated

We quantified the abundance of mRNAs, ribosome-associated RNAs, and proteins in Col-0 and Can-0 rosette leaves raised in growth chambers, under identical long-day conditions harvested at the same 9-leaf developmental stage (Additional file 4: Table S4). By minimizing environmental and temporal variation, we thereby focused on internal sources of variation in gene and protein expression and ribosome association. After quality control (see Methods), we identified 17,414 mRNA-expressed genes in common between Col-0 and Can-0.

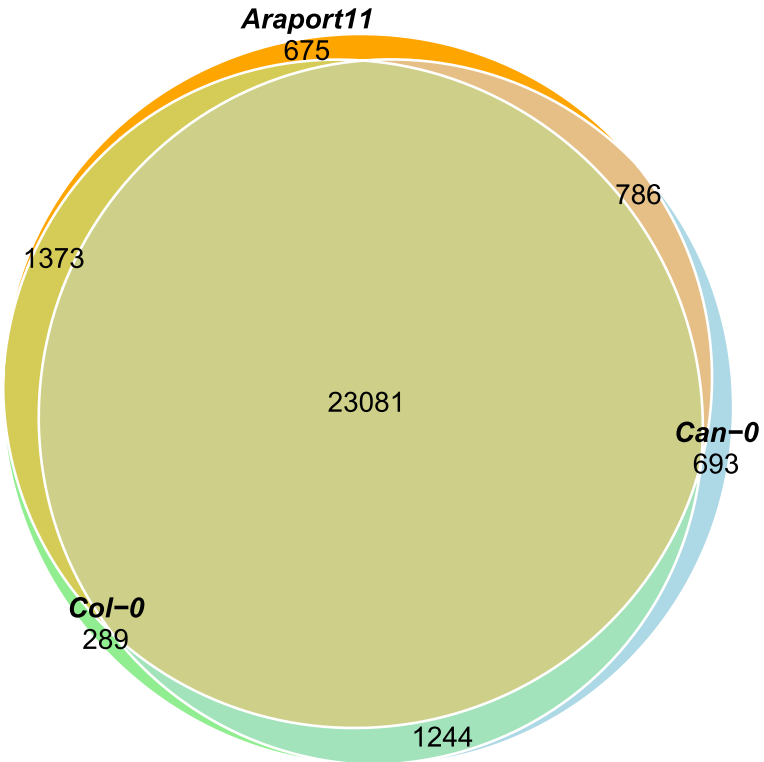

Fig. 3 Venn diagram showing overlaps between genes annotated in Col-0, Can-0, and Araport11. The numbers shown are the counts of genes in the different intersections. For example, there are 289 genes only found in the Col-0 annotation, while 1373 genes are common to Col-0 and Araport but absent from Can-0

Proteins were analyzed using a label-free data-independent acquisition (DIA) workflow, employing intensity-based absolute quantification (iBAQ; [22]) to derive comparative protein abundance from mass spectrometry data between the two accessions. This enabled quantification of 8915 proteins common to Col-0 and Can-0. Both mRNA and protein expression were detected in both accessions for 7771 ortholog pairs, while a further 9633 pairs had mRNA expression in both accessions but no observed protein expression. To some extent, this reflects the comparatively lower coverage of proteomic data, but it also suggestive of extensive post-transcriptional control. Interestingly, 28 genes had protein but no mRNA expression in Col-0 (and 32 in Can-0), perhaps indicating RNA instability. If we only retain genes with expression observed in all replicates, then there are 7721 genes expressed in mRNA and protein in both accessions and 7494 expressed in mRNA but not protein in both accessions; no genes were expressed in protein but not mRNA.

Both mRNA and protein measures were highly reproducible across biological replicates (Additional file 5: Table S5); for genes with both mRNA and protein expression, correlations of log-transformed mRNA levels between 5 replicates within an accession all exceed 0.98, as did correlations between 4 replicate protein levels within an accession (all correlations are between log-transformed data unless otherwise stated). Correlations between accessions are also very high; all mRNA replicates exceed 0.95, despite the presence of many differentially expressed genes (discussed later). The strength of these mRNA correlations is slightly lower for genes without protein, although all correlations between replicates still exceed 0.94. Correlation of protein expression between accessions always exceed 0.91 among replicates.

We then combined the expression levels across replicates for each gene by taking Geometric means. Looking across genes within each accession, the most abundantly expressed genes and proteins greatly exceed the respective median level: 2151-fold, Col-0; 1370-fold, Can-0 for mRNA and 640-fold; 591-fold, respectively, for protein (Fig. 4A, B). We report fold changes relative to the median rather than the full dynamic range of expression because the latter is strongly biased by genes expressed at near-zero levels. Protein dynamic range is well established to exceed that of RNA but is not generally captured in proteomics workflows, and especially not in tissues such as leaves which are dominated by a few highly abundant proteins [23, 24]. Most of the highly expressed genes and proteins are involved in photosynthesis, as would be expected for leaf tissue.

To analyze correlations within and between mRNA and protein expression, we focused on the 7771 ortholog pairs of genes with mRNA and protein expression in both Col-0 and Can-0. After log transformation, the correlation between mRNA and protein within an accession was (0.664, 0.647), about 50% higher than without log transformation (Table 1, Fig. 4A, B), and similar to that reported in maize leaves [1]. Despite differences in measurement and analysis methodologies, the most accurate predictor of protein expression is not mRNA from the same accession but rather protein from the other accession (Table 1). We return to this point below.

The distributions of mRNA levels are markedly different depending on whether protein expression is also observed or not (Fig. 4C, D), with the former following an approximately lognormal distribution while the latter has a complex mRNA distribution, comprising a spike of genes with very low expression and a shoulder of intermediate

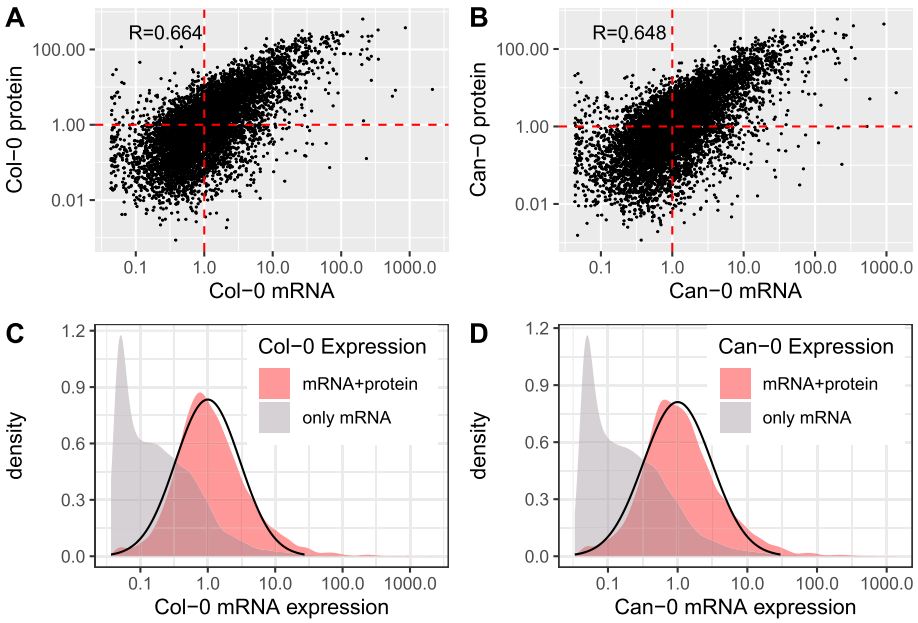

Fig. 4 The distribution of mRNA and protein expression, scaled so that the median level of expression of genes with both protein and mRNA expression is equal to 1. Scatter plots of mRNA (x-axis) vs protein (y-axis) expression for orthologous genes in Col-0 (**A**) and Can-0 (**B**). Dotted red lines show medians. Histograms of mRNA expression for genes with (pale red) or without (gray) detectable protein expression in Col-0 (**C**) and Can-0 (**D**). The black curves indicate lognormal densities fitted to the mRNA + protein histograms using robust estimates of mean and standard deviation. Expression scales are logarithmic throughout.

Table 1 Pearson correlations between mean expression levels of mRNA and protein in Col-0 and Can-0 across 7771 genes. Blue background: correlations between raw expression values. Orange background: correlations between log-transformed values, using the transformation $y = \log_{10}(x+1)$, i.e., adding a pseudo count of 1 unit

	mRNA Col	mRNA Can	protein Col	protein Can
mRNA Col	1	0.9722	0.6635	0.6382
mRNA Can	0.9679	1	0.6444	0.6471
protein Col	0.3101	0.3960	1	0.9154
protein Can	0.3104	0.4075	0.9406	1

expression, with a thin tail of highly expressed genes. The spike of genes with near-zero expression is due to those genes not being expressed in all replicates. Additional file 1: Fig. S3 shows the same distributions in Fig. 4C and D after omitting the $9633 - 7494 = 2139$ imperfectly replicated genes. It is possible that some of the imperfectly replicated genes represent artifacts, or are transiently expressed, or expressed in very rare cell types.

Isoform complexity does not affect mRNA-protein correlation

We next asked if alternative splicing affected the correlation between mRNA and protein. Although we could estimate the relative abundance of expressed mRNA isoforms for each gene from RNAseq data, it was not possible to estimate protein isoform

abundance in the same way, due to the sparsity of coverage of tryptic peptides. However, if isoforms of the same gene differ in their contribution to protein abundance, we reasoned that genes with a single isoform should exhibit a different (and likely higher) mRNA-protein correlation than genes with multiple isoforms. We therefore grouped the genes into subsets based on the numbers of their annotated isoforms and re-computed correlations within each subset. The forest plot in Additional file 1: Fig. S4 shows the mRNA-protein correlations for genes with up to 7 isoforms annotated in Col-0 and Can-0. The plot does not reveal any discernible trend between the correlation and the number of isoforms in either accession (the anomalous result for 7 isoforms likely represents sampling variation due to the few genes involved). We conclude that alternative splicing does not measurably impact the correlation between mRNA and protein levels.

Ribosome-associated transcript levels closely resemble standard mRNA expression for genes with protein expression

We next asked if transcripts associated with ribosomes were better correlated with protein expression. Ribosome-associated RNAs were quantified in six biological replicates each of Col-0 and Can-0 rosettes using 3′Ribo-seq [25] (expression levels are in Additional file 4: Table S4). In total, 17,513 of their 1–1 orthologs had detectable ribosome-associated transcripts (ribo-mRNA hereafter) in both accessions. Within the 7771 genes with protein expression, ribosome-associated transcripts behaved very similarly to the mRNA data described above; 7620 (97.9%) genes were also associated with ribosomes, and the correlation of log-transformed ribo-mRNA and mRNA expression levels was 0.9530 (Col-0), 0.9574 (Can-0). Correlation between the six biological replicates always exceeded 0.97 within an accession (and exceeded 0.94 between accessions, Additional file 5: Table S5). Their correlation with protein expression was (0.6490, 0.6415), very similar to that observed for mRNA determined by RNA-seq.

Within the 9673 genes for which mRNA but no protein was quantified, the pattern is slightly different. Of these, 6812 (70.4%) are ribosome associated, and the correlation of expression levels between mRNA and ribo-mRNA among these genes remained high, at (0.9303, 0.9451). All correlations between replicates within an accession exceeded 0.89 but between accessions, the correlations were lower with a minimum of 0.67. Only 59 genes are expressed in ribo-mRNA but absent from mRNA. Among the 29.6% of genes without ribo-mRNA expression, average mRNA expression was reduced by factors of (5.6, 5.4). In particular, the spikes of genes with very low expression seen in Fig. 4C and D are absent among ribosome-associated transcripts. Apart from this difference, the ribo-mRNA expression levels and patterns were essentially interchangeable with those of mRNA, and so for the remainder of this study we use the mRNA expression data.

Constitutive CpG methylation is reproducible across assay type and between accessions

We measured CpG methylation using both bisulfite-converted Illumina reads and ONT long reads, the latter collected as a by-product of generating sequence for de novo genome assembly. As Fig. 1A (orange track) shows, uniformly high levels (>75%) of CpG methylation occur throughout the centromeres but methylation is variable in other regions of chromosomes. Across CpG sites, the correlation between bisulfite and ONT gbM values was 0.94 in both Col-0 (2.6 M sites) and in Can-0 (2.8 M) despite the

different growth conditions (Additional file 1: Fig. S4). As the coverage of ONT data is superior to bisulfite sequencing and methylation readouts are less prone to GC bias [26], we therefore used ONT CpG methylation values in all further analyses.

We then computed gene-body methylation (gbM) for every annotated gene as the mean percentage methylation across all CpG dinucleotides within the genomic interval spanned by the gene, including exons and introns. We also quantified methylation in flanking regions around each gene for differential expression analysis (described later). The genome-wide distributions of gbM in Col-0 and Can-0 across gene expression categories (i.e., protein and mRNA, only mRNA, or no expression) are shown in Fig. 5. All three distributions share a mode in gbM around 12%, but with differing upper tails.

The correlation between gbM in Col-0 and Can-0 is slightly higher in genes with protein and mRNA detected ($R = 0.890$) than in genes with only mRNA ($R = 0.830$). Figure 4 also plots the distribution of (6195, 6042) other genes without any expression at mRNA or protein. These contain subsets of (838, 952) genes where gbM exceeds 70%, and which are concentrated in the shoulders of the highly methylated centromeres of each chromosome, i.e., matching the local CpG methylation levels (Fig. 5). Additional file 1: Fig. S6 shows the spatial distribution of GbM for those Col-0 genes without any expression, categorized by gbM. However, most genes in all three categories have low levels of gbM under 25%; all modes are close to 10%. Thus, constitutive gbM varies only slightly across most genes. We discuss its impact on expression later, after first considering codon composition effects.

Codon composition affects expression

Codon composition is known to affect both mRNA and protein expression [27–29]. We modeled its explanatory power in our data in order to establish its role in the information flow underlying the Central Dogma and to test the hypothesis that more frequent codons are associated with increased expression [30]. In our analysis, we represented the coding sequence of each protein-coding gene by a vector of the 61 non-terminator codon frequencies, hereafter abbreviated to CDS.

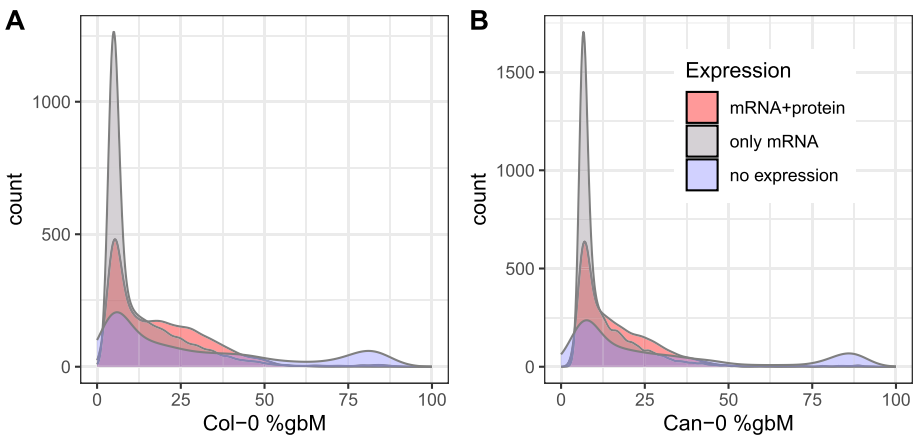

Fig. 5 Distribution of gbM in **A** Col-0 and **B** Can-0 genes, categorized according to whether both protein and mRNA were detected (pink) or only mRNA detected (gray), or not expressed (blue)

We modeled mRNA and protein expression of each gene in terms of these codon frequencies by fitting linear multiple regression models to the log-transformed expression levels, thereby estimating the expression effect (regression coefficient) of each codon. Under this model, if the codon c with multiple regression coefficient β_c occurs N_{gc} times in gene g, then its predicted log expression level is $y_g = \mu + \sum_c N_{gc}\beta_c$, where μ is the average expression level. Codons with positive codon effects increase expression and negative effects decrease it. The codon expression effects and their standard errors are shown in Additional file 6: Table S6.

We defined the standardized effect of each codon on expression as its expression effect divided by the standard error. Among 7771 ortholog pairs with mRNA and protein expression in both accessions, standardized effects are highly correlated between mRNA and protein in both accessions ($R = 0.800, 0.800$) (Fig. 6A, B, Additional file 1: Fig. S7). These codon expression effects are also highly reproducible between Col-0 and Can-0 ($R = 0.99$ for both mRNA and protein). If we estimate mRNA effects by restricting

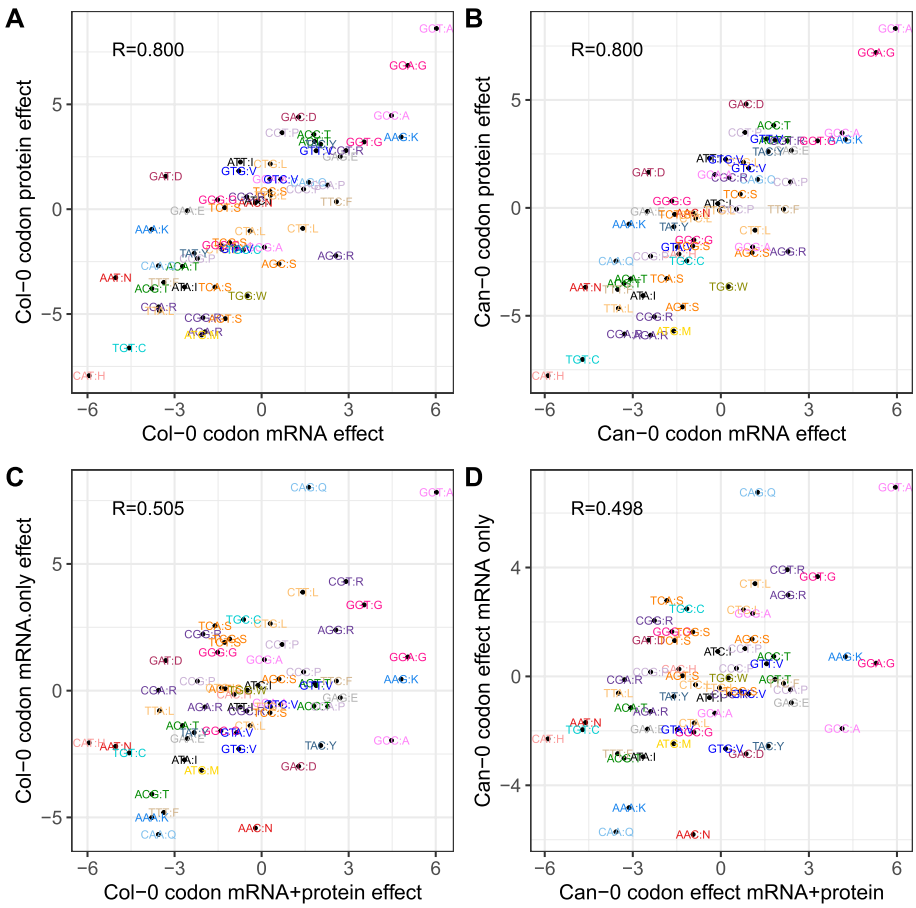

Fig. 6 Codon effects on mRNA and protein expression. **A** Scatter plot of 61 codon effects on log Col-0 mRNA expression (x-axis, represented as the *T*-statistic for each codon) vs the corresponding effects on log Col-0 protein expression. Each point is labeled with the codon and encoded amino acid, and all codons with the same amino acids share the same color. Models were fitted to genes with both protein and mRNA expression in Col-0. **B** Similar analysis for Can-0. **C** Similar analysis comparing codon effects on mRNA expression in Col-0 estimated for genes with both mRNA and protein expression (x-axis) and for genes with only mRNA expression (y-axis). **D** Same plot as for **C** but in Can-0

attention to genes for which protein was not quantified, then the resulting mRNA codon effects are markedly less correlated with those modeled from genes with both protein and mRNA ($R = 0.505, 0.498$, Fig. 6C, D).

We next asked if these codon expression effects are related to global codon frequencies. We computed global codon proportions, either by summing the genomic gene codon frequencies to give proportions independent of expression level, or by weighting the gene codon frequencies by mRNA or protein expression, thereby taking account of expression level. We then compared these proportions with the codon effects. The standardized codon effects for mRNA and protein expression are uncorrelated with overall codon abundance, defined as the relative fractions of codons across genes with both protein and mRNA expression (Fig. 7A, B; neither of the correlations of 0.065 and 0.235 are significant at $p < 0.05$). Thus, more commonly used codons are not associated with higher gene or protein expression.

We also tested if codon frequencies correlate with gbM by fitting the same multiple regression models with gbM as the dependent variable in place of mRNA or protein

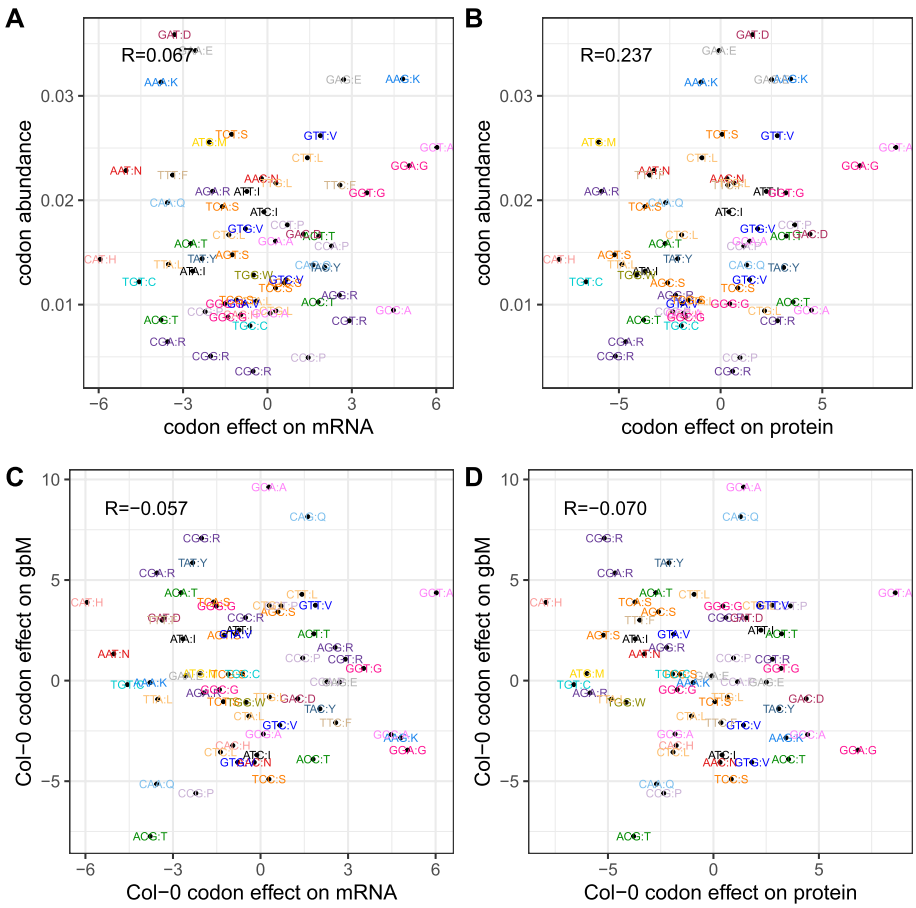

Fig. 7 Lack of correlation between standardized codon effects in Col-0 and gbM and codon abundance. In each scatter plot, each point represents a codon and is color-coded by the encoded amino acid. Standardized codon effects on gbM (y-axis) vs codon effects on mRNA expression (**A**) and protein expression (**B**). **C**, **D** y-axis is codon abundance, defined as the fraction of codons in all genes with mRNA and protein expression, x-axis: codon effect on mRNA expression estimated by multiple linear regression

expression. Figure 7C and D show that there is only weak correlation between the codon effects for expression and those for gbM. We return to the relationship between gbM and codons later.

tRNA abundance tracks global amino acid frequency

We next asked how tRNA abundance relates to codon expression effects, to codon abundance and gene expression. The genetic code is redundant, with 20 standard amino acids specified by 61 sense codons in eukaryotes. Isoacceptor tRNAs comprise families that accept the same amino acid, but which differ in their anticodon sequence, reflecting the fact that all amino acids other than methionine and tryptophan are specified by more than one codon. Isodecoder tRNAs carry the same anticodon but differ at their primary sequence at sites other than the anticodon. In common with other eukaryotes, *Arabidopsis* encodes tRNAs for only 45 sense codons plus the initiator tRNA-Methionine, the remainder employing third-base wobble base-pairing to effect translation [31–33] with specific tRNAs translating these missing codons [33] (Additional file 7: Table S7).

We measured tRNA abundance using modification-induced misincorporation tRNA sequencing (mim-tRNAseq) [34, 35] in Col-0 and Can-0 leaves grown and harvested under the same environmental conditions used to quantify mRNA and protein expression. Using the genomic tRNA database (GtRNAdb, [36]) annotation of *Arabidopsis* tRNA genes, we queried expression at 642 nuclear-encoded tRNA genes that also had Araport11 gene identifiers (Additional file 7: Table S7), representing 224 distinct tRNA isodecoders. We observed non-negligible expression for 157 of these isodecoders. For each of the 46 anticodons, we calculated the relative expression across all isodecoders.

Nuclear-encoded tRNA abundance levels are highly reproducible between Col-0 and Can-0 ($R = 0.988$; Fig. 8A). Within an accession, the relationship between codon frequencies and tRNA isodecoder abundance (Fig. 8B, $R = 0.397$) is obscured by the presence of codons without dedicated tRNAs (the vertically stacked codons on the left of Fig. 8B); some codons with no specific tRNA gene still have strong standardized effects. Figure 8C shows the correlation increases to 0.561 when we merge frequencies for codons translated by the same tRNAs. When we further aggregate tRNA abundance and overall codon usage by the encoded amino acid to produce isoacceptor frequencies, the correlation increases to 0.754 (Fig. 8D; results for Can-0 are very similar), and reaches 0.811 when amino acid frequencies are further weighted by protein abundance (a very similar relationship occurs when amino acid frequencies are weighted by mRNA abundance instead, because these are very highly correlated $R = 0.979$). Figure 8D shows the most discordant amino acid is tryptophan (W), which has higher tRNA expression than expected given its low frequency. Interestingly, tryptophan is the only amino acid apart from methionine encoded by a single codon. Additionally, it is encoded by UGG, which is in the same codon box as the UGA stop codon. High levels of its matching tRNA may be needed for it to compete with release factor, which may sample UGG codons.

In contrast to the strong relationship between tRNA and amino acid abundance, tRNA abundances are unrelated to codon expression effects. The correlations between tRNA abundances and the mRNA codon effects are (0.286, 0.283) (scatter plots in Additional file 1: Fig. S8 A, B) and correlations with protein codon effects are (0.249, 0.234) (scatter plots in Additional file 1: Fig. S8 C, D). These correlations are of only borderline

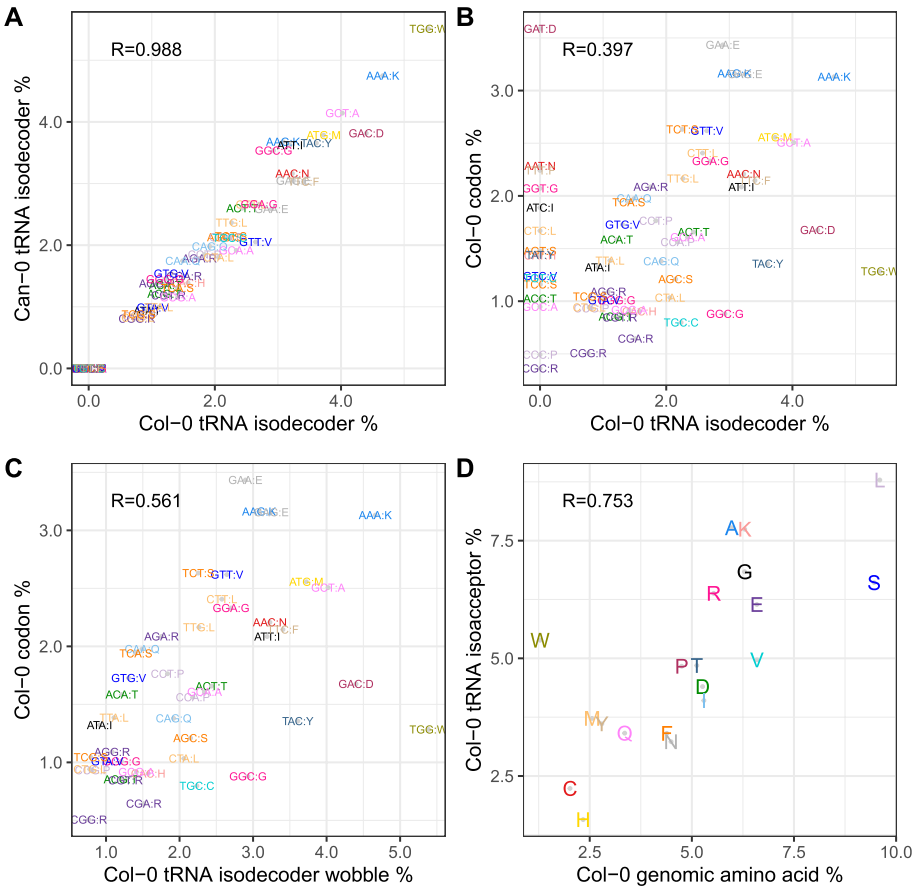

Fig. 8 Relationships between tRNA abundance and codon and amino acid frequencies. Codons and tRNAs are color-coded by encoded amino acid; the corresponding amino acid for each codon is specified in single letter code following a colon (i.e., in the format AAC:N; asparagine). Pearson correlation coefficients are shown in top left of each plot. **A** Col-0 tRNA isodecoder percentage (x-axis) vs Can-0 tRNA isodecoder percentage (y-axis). **B** Col-0 tRNA isodecoder percentage (x-axis) vs Col-0 codon frequency percentage. **C** Same as **B** except that codons without specific tRNAs are merged with the codons responsible for their translation to amino acids. **D** Col-0 codon fraction across all annotated genes (x-axis) vs Col-0 tRNA abundance aggregated by encoded amino acid (AA). Equivalent plots for Can-0 are very similar. If codon frequencies are re-weighted by the protein expression levels, the plot are very similar with slightly higher correlations

statistical significance (mRNA: $P < 0.025$, protein: $P < 0.057$). For comparison, the correlation of 0.811 with amino acid frequencies satisfies $P < 10^{-14}$. All codon-related statistics (regression coefficients and their standard errors, and tRNA abundance data) are in Additional file 7: Supplemental Table S7.

Modeling across omic levels reveals significant codon composition effects on expression and gbM

We next asked which genomic features predict mRNA and protein expression levels across genes within each accession. To model mRNA expression, the explanatory factors we considered were coding sequence DNA composition (referred to as CDS hereafter) and gene-body CpG DNA methylation (gbM, defined as the mean percentage of methylated CpGs within introns and exons). For protein gene expression, we additionally

considered mRNA expression as an explanatory factor. We also modeled gbM in terms of CDS. Taken together, these models encapsulate the Central Dogma's information flow.

We fitted multiple linear regression models, where the focal dependent variable could be gbM, mRNA, or protein, across various subsets of expressed genes, and the independent variables the lower omic levels measured in the same genes. For this modeling, to investigate the hypothesis that mRNA levels capture all the information in the underlying DNA sequence relevant to protein expression, we reparametrized CDS codon frequencies as the combination of three, biologically interpretable, nested components of increasing complexity, namely protein length (the sum of all codon frequencies), then the 20 amino acid frequencies (the Sums of frequencies for those codons representing a given amino acid, requiring 19 additional parameters), and finally codon usage within each amino acid (representing the deviations from the baseline effect for an amino acid, comprising 41 extra parameters). The first two of these components are linear combinations of the codon frequency counts, and the third accounts for variation due to the choice of codon within an amino acid class. For clarity, we refer to "codon frequencies" when interpreting CDS as a model of the 61 codon counts (i.e., as in the previous sections) and "codon usage" when interpreting it after separating out protein length and amino acid frequency.

Thus, depending on the parameterization used, the same model can be reinterpreted to provide different insights, while yielding identical predicted effects and explaining the same total variance. This reparameterization was used to test the effects of adding increasing information about sequence composition on constitutive gbM, mRNA, or protein expression within a single analysis of variance.

Each model is expressed in the form $Y \sim A + B + \ldots$, where Y is the target omic level and A, B … are explanatory omic levels. For example, the model $Col\text{-}protein \sim CDS + gbM + mRNA$ means that protein expression across the genes with protein and mRNA expression in Col-0 is modeled in terms of first codon frequencies (CDS), then gene-body methylation (gbM) and finally mRNA expression measured in those genes. The order in which levels are included in a model affects how much variation is explained by each level (i.e., the modeling is greedy, so each level is assigned the maximum possible variation after allowing for the previously fitted levels) thereby revealing confounding between levels. For example, fitting gbM either alone or after fitting CDS reveals how much variation in gene expression is solely attributable to gbM. We also subdivided the genes into two classes; those 7771 with both measurable mRNA and protein expression—the difference is due to a few genes with multiple annotated stop codons which were excluded—and those 9633 with only mRNA expression and denoted mRNA* or gbM* in Fig. 9 and Additional file 8: Table S8. Both mRNA and protein expression were log-transformed prior to fitting the linear models which therefore represent multiplicative effects on expression. Figure 9 legend describes the dependent and explanatory omic levels in detail.

The analyses are summarized as bar plots in Fig. 9, where the horizontal extent of each bar indicates the fraction of variance in the target omic level attributable to the corresponding component in the model, after fitting the preceding terms to the left in the bar. For comparison, we also show the results of modeling an omic level in one accession by the corresponding level in the alternate accession (tan bars: gbM, pink bars: mRNA, and

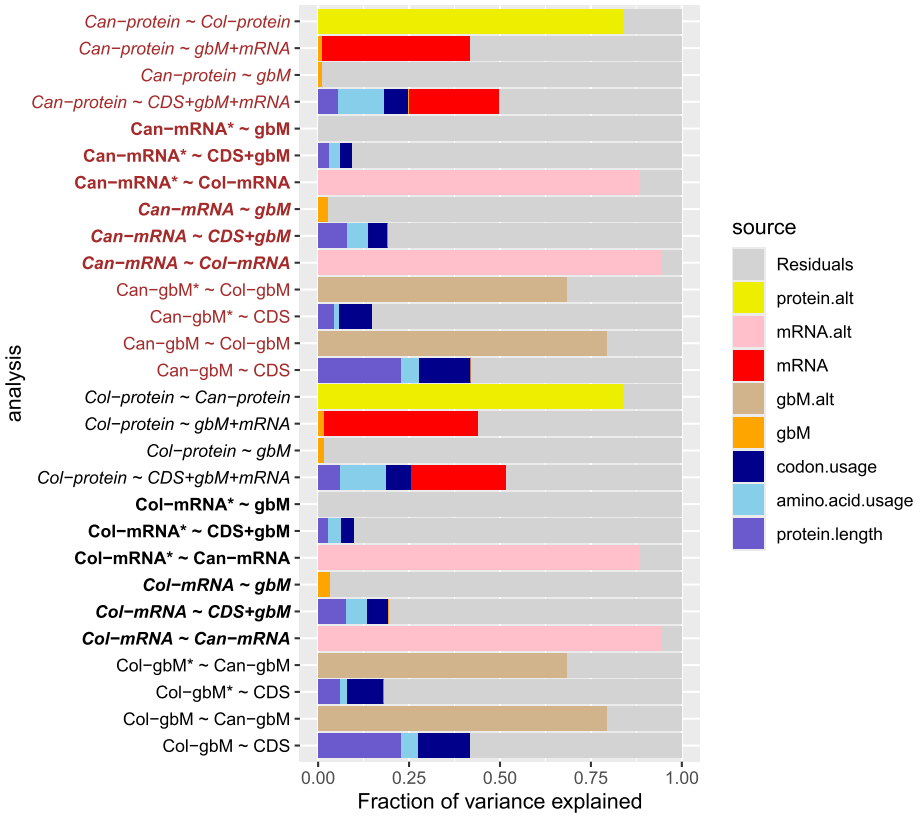

Fig. 9 Bar plots of variance explained by multiple linear regression models. Each row represents one model. The model is specified on the left, the color indicating whether Col-0 (black) or Can-0 (brown) is the target omic level (dependent variable). Each model is represented by a formula $Y \sim A + B + \ldots$, where Y is the target and A, B … are explanatory omic levels. The targets for the Col-0 analyses are as follows: Col-protein: log-transformed protein expression, Col-mRNA: log-transformed mRNA expression for genes also with protein expression; Col-mRNA*: log-transformed mRNA expression for genes without protein expression; Col-gbM: percent gene-body methylation for genes also with protein expression; Col-gbM*: percent gene-body methylation for genes without protein expression. Similar names apply for the Can-0 analyses. The explanatory omic levels are as follows: CDS: coding DNA sequence composition (partitioned into protein length, amino acid usage, and codon usage); gbM: percent gene-body methylation; mRNA: log-transformed mRNA expression; gbM.alt, mRNA.alt, protein.alt: expression of gbM/mRNA/protein in the alternative accession (i.e., Can-0 if the target accession is Col-0). The bar plot for each analysis represented the fraction of variance explained by each term in the model, using the color-coding given in the legend. CDS effects are partitioned into protein.length, amino.acid.usage, and codon.usage; the horizontal extent of each bar represents the fraction of variance due to the corresponding variable, after first fitting the preceding variables in the formula from left to right

yellow bars: protein). The results for Col-0 (Fig. 9 upper) and Can-0 (Fig. 9 lower) are extremely similar, illustrating the robustness of our results to genetic perturbation.

The p value of each variance component from its corresponding partial F-test is given in Additional file 8: Table S8; virtually all components are extremely significant with analysis of variance p values often much smaller than 10^{-10} even when the fraction of variance explained is too small to be visible. That is, statistical significance is necessary but not sufficient to imply biological importance. Multiple linear regression models are "greedy": the order in which explanatory variables are fitted in the model determines how much variance each explains.

We consider the impact of explanatory omic levels on mRNA and protein expression in Central Dogma order. The simplest explanatory variable, protein length, is known to

be anticorrelated with expression [37]. In Col-0, we find protein length alone explains 7.7% of the variation in mRNA expression (double that explained by gbM at 3.5%) and 6% of protein expression variation. The statistics in Can-0 are mRNA: 8.1% and protein: 11.1%. The next component of sequence composition, amino acid usage, explains significant additional variance in mRNA (5.7%, 5.8%) and in protein expression (12.8%, 12.7%), after accounting for sequence length, although spread across 19 parameters. Codon usage explains further variance (mRNA:6.0%, 5.3%; protein: 6.8%, 6.7%) but spread over far more (41) parameters. Overall, CDS effects on mRNA expression (19.5%, 19.2%), are lower than protein expression (25.6%, 24.9%), and the relative impacts of the three CDS components also differ. When we model mRNA expression of just genes without protein expression (denoted mRNA* in Fig. 9), the fractions of variance explained by CDS are halved (10.1%, 9.4%). The impacts of mRNA alone on protein level variation are found by squaring the correlations in Table 1 and Fig. 4, yielding (44.1%, 41.9%), which are also almost identical to the combined effects of gbM and mRNA (Additional file 8: Table S8), showing that gbM effects on protein expression are also mediated by mRNA levels.

We then modeled constitutive gbM as a function of CDS. Among genes with both mRNA and protein expression, we find CDS effects account for close to half (42.0%, 42.0%) of gbM variance. However, these fractions are more than halved (18.1%, 14.8%) among genes with only mRNA expression. Modeling gbM in one accession by the corresponding level in the alternative accession (shown as tan colored bars in Fig. 9) explains far more of the variance (79.4% for genes with protein and 68.6% for those without) confirming constitutive gbM is highly reproducible (as expected, given the reproducibility of the underlying CpG methylations reported above) but that 58% of this variation is unexplained by codon frequencies.

We next treated gbM as an explanatory level to model mRNA and protein expression. When considered in isolation (i.e., excluding CDS effects), gbM has a relatively small but nonetheless statistically significant impact on mRNA expression in genes with both protein and mRNA expression, explaining (3.5%, 4.2%) of mRNA variance, but a negligible impact on the expression of genes without protein (0.1%, 0.1%). When CDS effects are fitted before considering gbM, virtually all the effects of gbM are ablated; under a constant environment, the impact of constitutive gbM on mRNA expression mediates a small fraction of sequence composition effects.

If mRNA expression is added after CDS and gbM, in total half of the variation in protein expression can be explained (51.9%, 49.8%), and which exceeds that when only gbM and mRNA are included (44.1%, 42.0%). Thus, part of the information encoded in CDS relevant to protein expression is not mediated through mRNA, in contradiction to the Central Dogma. Both CDS and mRNA account for non-negligible and independent fractions of protein expression variance. While there is considerable confounding between mRNA and CDS (because CDS explains much mRNA variance), it is clear that, at minimum, mRNA expression explains an additional $51.7-25.7=26.0\%$ of Col-0 variance that cannot be accounted for by CDS and gbM, and CDS explains, at minimum, an additional $51.7-44.1=7.6\%$ of Col-0 variance that cannot be explained by mRNA and gbM.

Similarly to modeling gbM, much greater fractions of variation are explained by modeling Col-0 mRNA by Can-0 mRNA (and vice versa) (94.5%, 94.5%) or Col-0 protein by Can-0 protein (83.8%, 83.8%). Thus, while our simple regression models are powerful,

they do not capture all the information, and there is unexplained yet reproducible variation.

Comparisons between accessions

Our analysis has so far has emphasized the strong similarities between Col-0 and Can-0, despite their ~1% genetic difference. In this section, we ask how differential mRNA or protein expression between the accessions relates to differences in CDS and to CpG methylation and other features.

Sequence differences between orthologs reduce expression correlations

We first investigated if the high correlations observed between the same omic level measured in Col-0 and Can-0 (Fig. 9) are affected by sequence differences. Among the 24,325 annotated ortholog pairs of protein-coding genes, the median protein lengths are (349, 348) amino acids, and the median difference in codon frequencies between orthologous genes in Col-0 and Can-0 (i.e., the Sum of absolute differences in the counts of each of the 61 non-terminator codons) is 6 codons. Additional file 1: Fig. S9 shows the inter-accession mRNA-mRNA and protein–protein correlations when the genes with both protein and mRNA expression in both accessions are grouped according to their codon frequency differences. The plot shows that there is a slow but steady reduction in mRNA-mRNA correlation as the numbers of codon differences increases, but that the protein–protein correlation reduces dramatically when there are over 50 codon differences.

Differentially expressed genes and proteins

There are 7,585 differentially expressed (DE) mRNAs at FDR < 0.05 among the 17,771 orthologous Col-0:Can-0 gene pairs at which we could make a determination using EdgeR, ignoring protein expression status. In the subset of 7060 genes with differential determinations for both mRNA and protein, we observed 866 DE proteins (FDR < 0.05) and 2850 DE mRNAs (FDR < 0.05; EdgeR did not determine DE status for all genes, so these subsets are slightly smaller than those in the previous sections) (Additional file 9: Table S9). Figure 10A plots the $\log_2$ fold change in expression (logFC) for mRNA vs protein, color-coded by DE FDR. It shows that logFC is broadly consistent between mRNA and protein, although there are many genes which are DE for only mRNA or only protein. Of the DE mRNAs and DE proteins, 579 (20% of DE mRNAs and 67% of DE proteins) are in common (Fisher's exact test < 10^{-32}), as shown in the Venn diagram

(See figure on next page.)

Fig. 10 Differential expression (DE) and differential gene-body methylation (gb DML). **A** Scatter plots of $\log_2$ fold change (logFC) for mRNA (x-axis) vs protein (y-axis) for 5698 ortholog pairs between Col-0 and Can-0 with DE determinations in protein and mRNA. Points are color-coded according to whether the pairs are DE at both protein and mRNA, only protein, only mRNA, or neither (all determinations using FDR < 0.05). **B** Venn diagram of overlaps of mRNA and protein DE and gb DML. Numbers are counts of DE gene pairs within each subset (e.g., there are 2026 pairs that are only DE mRNA, 524 + 55 = 579 that are both DE mRNA and protein, and 55 that are DE mRNA and protein and gb DML). **C** Distributions of absolute log2 fold change in mRNA expression between Col-0 and Can-0, for genes with (orange) or without (green) gb DML. In **A** and **C**, the figures are truncated to omit the small fractions of absolute log2 values exceeding 3.0

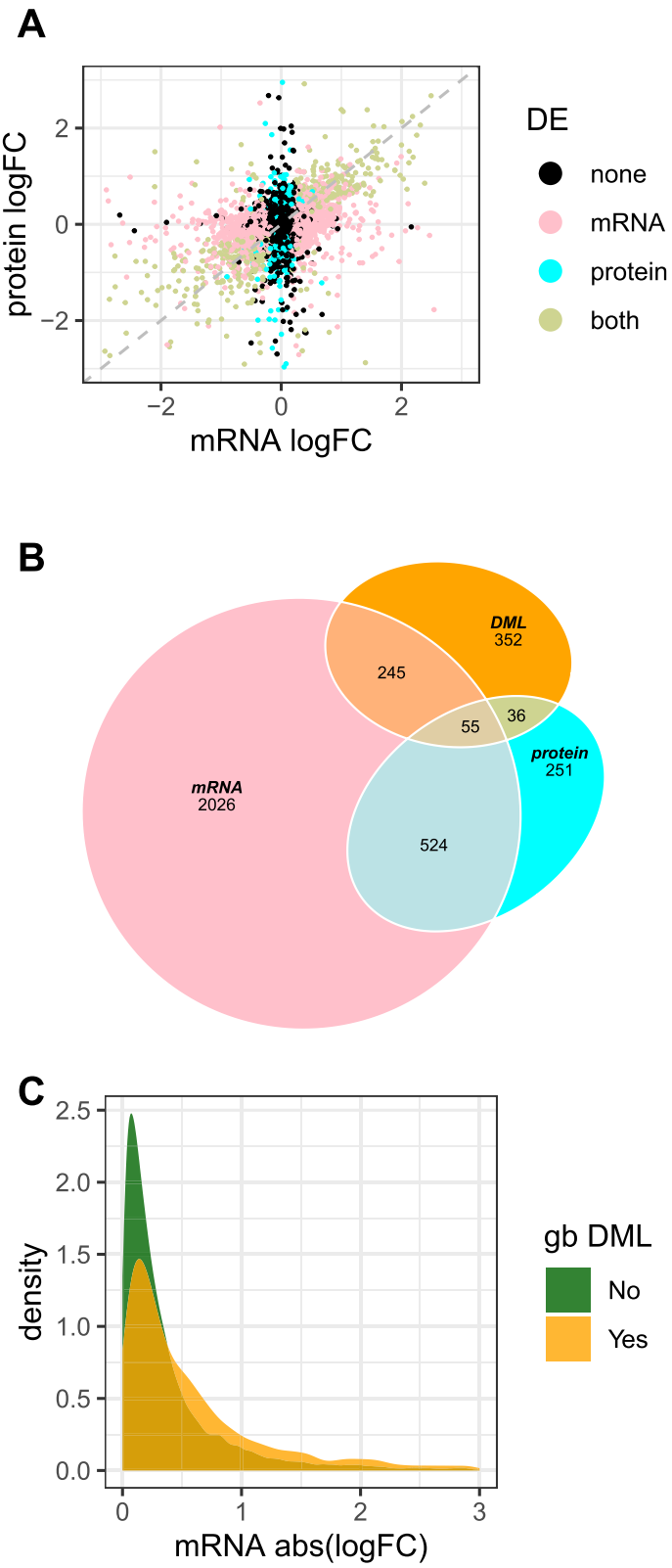

Fig. 10 (See legend on previous page.)

in Fig. 10B. Gene Ontology (GO) enrichment analysis [38] reveals distinct but overlapping enrichments between DEGs (Additional file 1: Fig. S10) and DEPs (Additional file 1: Fig. S11). Proteins and transcripts associated with glucosinolate biosynthesis exhibited increased abundance in Can-0, whereas proteins associated with immunity ("hypersensitive response," "cell death," "response to biotic stimulus") and abscission were increased in abundance in Col-0.

Genetic and epigenetic correlates of differential expression

We distinguish between a differentially methylated locus, DML—a single syntenic CpG dinucleotide at which methylation differs between Col-0 and Can-0—and a differentially methylated region (DMR), in which average methylation differs across the CpG dinucleotides within the region. If a gene body overlaps with at least one DML or DMR, this gene body is defined to be also DML or DMR. We used the same definition to classify intron and exon regions, and genomic contexts up- or downstream of the gene body, and for structural variants (SV) as discussed below. The Venn diagram in Fig. 10B shows the overlaps with gb DML genes at 5% FDR. Of the 688 genes with gb DML, 336 (49%) are DE for mRNA or protein, or both.

We compared DML and DMR gene classifications with the corresponding absolute values of logFC expression, reporting $-\log_{10} p$ values (logP) of the Mann–Whitney tests, which are robust non-parametric tests of differences in the average ranks of the absolute logFC of expression between genes with or without differential methylation. This analysis therefore does not require differential transcript or protein expression to exceed any specific FDR threshold but instead considers trends. The results are summarized in Fig. 11 (Additional file 9: Table S9).

Despite the high gbM correlation between Col-0 and Can-0 at orthologous genes ($R = 0.89$), the presence of DML or DMR within or nearby a gene body strongly associates with absolute log-fold changes in mRNA expression (Fig. 11). DML and DMR inside gene bodies, or within 100 bp upstream, have the highest impact on differential mRNA expression. In general, the presence of even a single CpG methylation difference (i.e., DML) is a stronger predictor of gene expression difference than is DMR, apart from intronic DML which is not significant for mRNA abundance (logP= 0.61), but highly significant for intronic DMR (logP= 122). Figure 10B shows the overlaps between DE mRNAs and proteins and gb DML. We found that the corresponding signed Mann–Whitney tests (i.e., where we did not take absolute values of logFC) were markedly less significant. Figure 10C shows the distributions of absolute log-fold changes for mRNA expression for genes with or without gb DML. Although the distributions appear broadly similar, they have highly a significantly different Mann–Whitney statistic (logP = 79). Thus, although differential methylation is strongly statistically associated with differential expression, it does not reliably predict the direction of change.

We then used the same methodology to test if sequence differences are associated with absolute differences in mRNA and protein expression. We tested for associations between the presence of SNPs or small (< 10 bp) indels, structural variations (SV, defined as indels > 10 bp), or nearby transposable elements (TE). Gene-body SVs are strongly associated with differential mRNA abundance (Fig. 11A)—intronic SVs have the second strongest association with differential mRNA expression of any

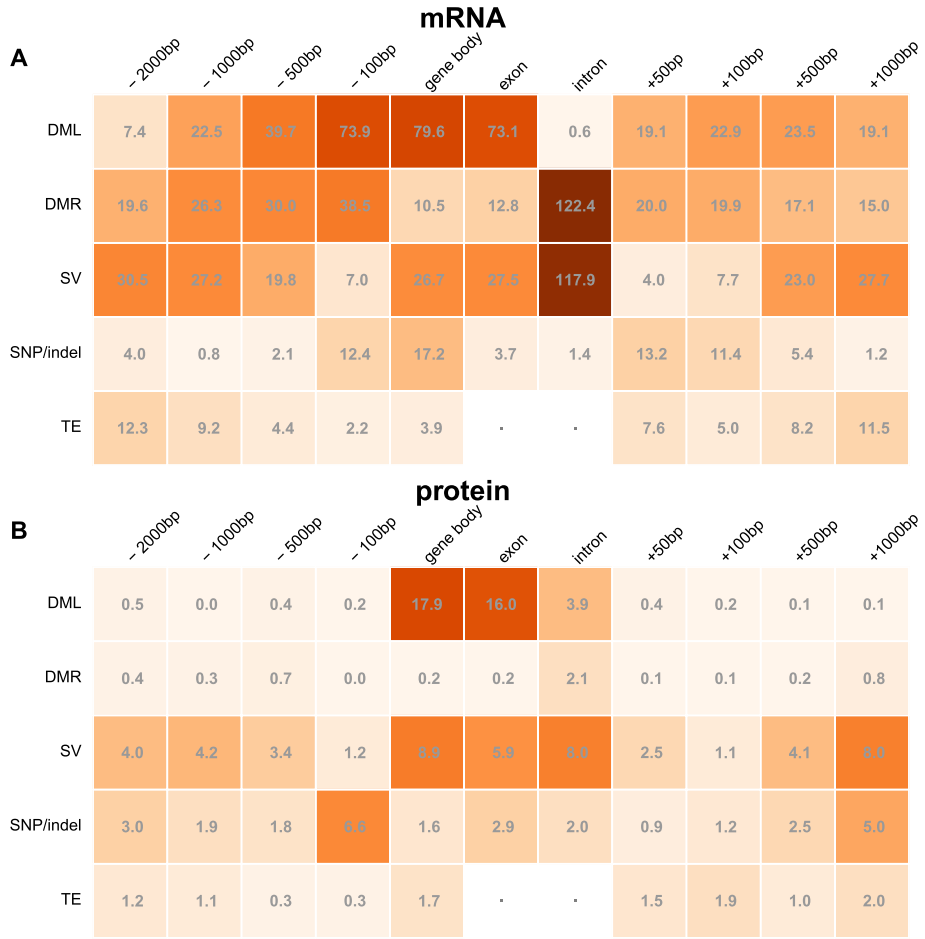

Fig. 11 Impact of differential methylation, structural variation, and indels on differential (**A**) mRNA and (**B**) protein expression between Col-0 and Can-0. The x-axis represents a schematic gene, comprising upstream, gene-body (subdivided into intronic and exonic components) and downstream genomic contexts. The y-axis represents the variation categories DML: differentially methylated loci, DMR: differentially methylated regions, SV: structural variations, indels: short insertion-deletions, and TE: transposable elements. The number in each is the negative $\log_{10} p$ value of the unsigned Mann–Whitney test of association between the category in the given genomic context and differential expression (except for TE which shows the logP of the Spearman rank correlation test reflecting the fact that differential TE abundance is measured quantitatively). The orange shade of the cell indicates the strength of association, from dark (strong) to pale (weak)

feature tested, logP = 117, only slightly less than that for intronic DMR (logP = 122); this is because the presence of intronic SVs and DMR frequently co-occur (contingency table odds ratio of 62.2, logP = 1873), suggesting intronic SVs cause intronic DMR. It is unclear if these SVs alter expression directly or are mediated by modulating methylation.

Interestingly, distant structural variants also have strong associations with differential gene expression. Differences in TE between the accessions have modest associations with changes in mRNA expression and again, absolute differences are more strongly correlated than signed differences, and are most strongly associated when they are 500–1000 bp upstream or downstream of the gene (maximum logP = 12.3). Differential TE seem to be uncorrelated with absolute differential protein expression, however.

Discussion

It is a remarkable fact that although most genes are encoded only once in the nuclear genome, their constitutive expression levels vary by orders of magnitude [24]. In the *Arabidopsis thaliana* rosette leaves studied here, the most highly expressed genes exceed the median level by over 1,000-fold for mRNA and over 500-fold for protein. The key questions address these levels' reproducibility, and how they are set and maintained. We have shown here that in two genetically divergent accessions of *Arabidopsis thaliana*, these levels are indeed highly reproducible between biological replicates grown in a controlled environment, and that a simple multiple linear regression model based on gene codon frequencies is unexpectedly powerful at modeling both mRNA and protein expression levels. A gene's constitutive expression is thus partially determined by its internal codon composition, which presumably evolved to express it at its optimal level by selecting codons adaptively and tuning tRNA abundance to match overall demand for amino acids. The diagram in Fig. 12 summarizes our findings as a flow chart showing the relative importance of different omic levels to expression.

Codon and tRNA effects on expression

Codon frequencies alone account for about 19% of the variance of mRNA expression and about 25% of protein expression. Augmenting the codon model of protein expression with mRNA expression levels almost doubles the variance explained to about 46%. Interestingly codon frequencies only explain 9% of variance among genes with mRNA but no protein expression, suggesting their expression is controlled differently.

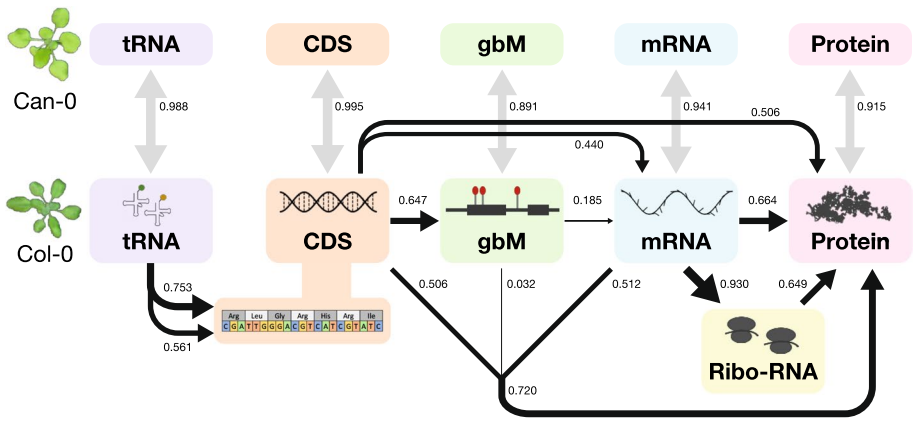

Fig. 12 Genomic information pathways and their numerical linkages as observed in this study, in relation to the Central Dogma. The boxes show different omic levels in Col-0 and Can-0. Black arrows indicate the flow of information implied by the Central Dogma starting from CDS (peach) via gbM (green) to mRNA (light blue), ribosome-associated mRNA (yellow) and finally protein (pink). The strength of Pearson correlation between levels is shown both by the numbers and by the thickness of the corresponding black arrows and represents the correlation between quantitative expression levels (except in the case of CDS where is represents the correlation derived from the multiple regression codon frequency model, i.e., the square root of the fraction of variance explained). The merged arrows connecting CDS, gbM, and mRNA to protein show the result of combining information into a single model; note that their combined correlation of 0.720 is less than the sum of their individual effects. The correlation of tRNA levels (lilac) with genome-wide codon and amino acid frequencies is shown on the left. The correlations between corresponding Col-0 and Can-0 omic levels are shown next to the gray double-headed arrows

When interpreting these results, it helps to bear in mind that the total fraction of variance explained by a model equals the squared correlation between the observed and fitted values. That is, correlations are larger than their equivalent variance fractions; 19% variance is equivalent to a correlation of 0.46. In Fig. 12, all the numerical linkages between omic levels are shown on the correlation scale. However, it is more meaningful to report variance fractions when decomposing a multi-component model in an analysis of variance (Fig. 9). In addition, we report results after log-transforming expression, so our models are multiplicative on the original measurement scale. The impact on the original scale of expression of the codon c with gene frequency n will be proportional to β_c^n, where β_c is both the regression coefficient we estimate and the multiplicative impact of a single copy of the codon on the original measurement scale.

If full sequence information from each gene is used for prediction (i.e., including the order of bases), instead of being Summarized as the 61 codon frequencies as was done here, it is possible to train large language models with millions of parameters to predict expression patterns and genomic annotations [39, 40]. In a different *Arabidopsis* data set, and encoding each gene sequence by its sequence of codons, mRNA prediction accuracies of between $R^2 = 0.2 - 0.4$ were achieved, depending on the model used [7]. Thus, there is additional, exploitable, information encoded in the order of the codons in each gene. Nonetheless, it is remarkable how well Our simple, biologically interpretable, model performs. We find that a given codon has similar impacts on mRNA and protein expression; the correlations between the 61 non-terminator codon effects on mRNA vs protein expression ($R = 0.80$, Fig. 6A, B) exceed those of the actual expression measures ($R = 0.64 - 0.66$, Fig. 4A, B), suggesting that an underlying biological mechanism is being isolated.

While codon effects could point towards tRNAs as mediators for their effects on protein translation, our analyses support the hypothesis that codons which increase protein translation also reduce mRNA decay, driving mRNA stability and hence mRNA abundance [41–43]. Interestingly, we find that tRNA abundance is uncorrelated with codon expression effects but instead tracks overall codon abundance (Fig. 8B, C). Furthermore, the aggregated tRNA abundance across all tRNAs specifying a given amino acid (i.e., isoacceptors) tracks the overall frequency of that amino acid across the proteome even more strongly than at the codon level (Fig. 8D). Our results suggest that tRNA abundance adapts dynamically to the overall demand for amino acids but that differences in translation efficiencies between tRNAs do not measurably affect abundance of specific genes. In summary, codons interact with two distinct phenomena—their frequency relates to tRNA abundance, but independently of their impact on expression.

The correlations we observed between tRNA abundance and codon usage resemble those reported in humans [33]. In addition, our codon expression effects resemble those from studies in human cell lines [44], where a combination of sequence features (including protein length and predicted mRNA decay rate) and mRNA expression explained about two thirds of the variance in expression of 512 proteins. Another study of mRNA half-lives in human and mouse [45] also identified codon frequencies and protein length as key factors. Although we did not measure mRNA decay rates, it is likely that Our models of mRNA levels are implicitly modeling them. In fact, of the 18 codons impacting mRNA half-lives in humans in [37], Our data in Additional file 1: Fig. S7 share the same

sign in at least 16 cases, a statistically significant coincidence ($P < 0.008$, see Additional file 11: Text T1), and suggesting these codon effects may be evolutionarily conserved.

Our data were all collected under uniform unstressed conditions, so we could not measure the impact of stress on codon usage, which affects the translation of specific codons [46]. Ideally, this study should be repeated across more genomes, tissues, and environments to take account of expression quantitative trait loci, cell-type effects, and environmental variation. The two genomes used here are from the set of 19 founders of the *Arabidopsis* MAGIC population of recombinant inbred lines descended from these founders [14, 47]. Work is underway by our group to analyze omic levels across the founders and the MAGIC population, which will enable us to test if our conclusions hold in the presence of significantly more genetic variation.

Our codon model is mathematically equivalent to the combination of protein length, amino acid frequency, and codon usage. Once refactored in this way, we find mRNA expression decreases with gene length, and that the choice of encoded amino acids and the choice of codon each have significant but smaller impacts than protein length. The relative impacts of these factors on protein expression are subtly different: for example, gene length is less important than amino acid choice. In addition, among genes with mRNA but without protein expression the impacts of all three factors are much reduced (Fig. 9).

It is noteworthy that a codon's impact on expression is entirely unrelated to its genome-wide frequency among expressed genes; it is not true that highly expressed genes preferentially use high frequency codons. Therefore, to increase the expression of a protein by editing its codon composition, one should select optimal codons based on their effects estimated from a model like that fitted here but derived from expression data from the species of interest. If the close similarity of codon effects on mRNA and protein expression that we observed generalizes to other species, then it might be sufficient to train codon models on mRNA expression alone, which is experimentally more tractable, and then extrapolate to protein expression. However, there will be important nuanced effects of codon usage for individual genes. Synonymous substitutions can influence diverse mechanisms related to gene expression and protein homeostasis including transcriptional regulation, mRNA lifetime, translation initiation efficiency, translation elongation rate, and downstream effects on protein folding and degradation [29].

Constitutive gene-body methylation effects

When the environment is controlled, our results minimize the role of constitutive gene-body methylation in regulating expression [12], despite its high conservation between accessions. Augmenting codon models of mRNA or protein expression with constitutive gbM makes only a negligible improvement. Indeed, constitutive gbM itself is largely predicted by local codon frequencies, explaining about 44% of gbM variation under unstressed environmental conditions. At most, constitutive gbM mediates a small fraction of the information in codon frequencies—explaining about 3–4% of mRNA and protein variation in the absence of codon frequency data—but does not contribute new information. Since constitutive gbM is highly conserved between accessions, it is indeed plausible that it is driven by local sequence context.

Many codons with strong effects on gbM do not contain CpG dinucleotides. This suggests that non-CpG local sequence context drives CpG gbM. Additionally, while protein length is the major determinant of gbM variation in genes that exhibit both protein and mRNA expression, it is irrelevant among genes with only mRNA expression, illustrating major differences in the genetic architecture of gbM depending on protein expression.

However, we caution against the view that constitutive gbM is irrelevant to expression. First, the distribution of gbM is tightly concentrated around 10% (Fig. 5A, B) with little variation and hence limited opportunity to influence expression. However, there is a distinct subset of several hundred centromere-associated genes without any mRNA or protein expression in rosettes for which gbM exceeds 75%. The expression of these genes—concentrated in the shoulders of the centromeres—might be actively silenced by these high gbM levels, but this does not explain how a further ~5000 genes are neither expressed nor highly methylated, and indeed highly methylated genes may simply be passively reflecting high methylation levels around the centromeres.

Second, differential methylation, both in gene bodies and elsewhere, is strongly associated with differential mRNA expression between Col-0 and Can-0. The presence of a single differentially methylated CpG (DML) is generally a better indicator of DE than is the average difference of methylation (DMR), but the direction of the change in expression is uncertain; unsigned association tests which ignore the direction of the DE are more significant than signed tests. Thus, DML are markers of perturbations in methylation due to local sequence differences, which increase the variance in expression rather than its direction (Fig. 10C). Differential methylation outside of gene bodies influences mRNA expression less than does gbM DML and has markedly less impact on protein expression (Fig. 11). Lastly, differences in transposable elements between the accessions are very weakly associated with differential expression of nearby genes (Fig. 11).

Genes that express protein and mRNA differ from those that only express mRNA

The distribution and causal modeling of mRNA expression is very different for genes which also exhibit protein expression, compared to those for which protein was not detected in our proteomics workflow. It is not true that moderately high mRNA expression necessarily implies any protein expression (Fig. 4). In fact, many genes without protein expression have higher mRNA levels than those with protein. This discordance has been noted in other species [1–6]. Those genes expressing both mRNA and protein have an overall distribution close to lognormal. Taken together with the log transformations used here, we suggest that these genes expression levels result from a balance between independent multiplicative processes of synthesis and decay. In contrast, genes without protein expression have a more complex distribution.

Surprisingly, using ribosome-associated mRNA expression levels does not change the picture, at least in regard to genes with both protein and mRNA expression. The only noteworthy difference appears in genes with mRNA expression but not protein—about 30% of which do not appear to be associated with the ribosome. However, it should be noted that 3′RiboSeq merely quantifies transcripts associated with ribosomes, without distinguishing between monosomes and polysomes and does not indicate whether the transcripts are actively translated. It may be that stronger correlations with protein expression translatome data could be obtained from ribosome profiling in which it is

possible to accurately determine the average number of ribosomes per mRNA and thus estimate the relative translation levels of a transcript [48].

Omic differences between genomes

Our study emphasizes the similarities between accessions. However, our analysis of their differences reveals some unexpected results. In particular, differential methylation—as a result of sequence differences—is a marker for differential expression, although the direction of the effect is unpredictable—it seems to act by increasing the variance of expression (Fig. 10C). Further modeling is required to turn this observation into a predictor of how sequence differences translate into omic effects.

We observed more genes with differential mRNA than protein expression. This may reflect the limitations of quantitative proteomics which do not readily permit sampling of the entire proteome, but it may also indicate buffering of the proteome. The discrepancy may also be related to the fact that protein expression is estimated (here) from peptide DIA data and is therefore subject to different measurement issues and different algorithmic processing steps than RNAseq data. If this step introduces biases, they are likely to be the same within genes across replicates, thereby contributing to differences between genes, and consequently mRNA/protein correlations. Even with these experimental and methodological issues, it is remarkable how closely estimated codon mRNA expression effects resemble protein effects (Fig. 6).

The best predictor of expression in one accession is expression in a different accession

Despite the success of the codon models in Fig. 9, the expression of gbM, mRNA, or protein in a given focal accession is more strongly correlated with orthologous expression in the other, genetically distant, accession than with lower omic levels within the focal accession. This orthologous expression fidelity remains unexplained. Possibly this is due to active homeostatic feedback mechanisms and the conserved effects of transcription factors on the control of expression under unstressed conditions. In a future study, it will be interesting to determine to what extent this predictive power is maintained in the presence of a stimulus such as a stress or a developmental cue. Another interesting question is what underlies the residual variation in constitutive protein and mRNA expression that is unexplained by the models used here. This is not due to measurement noise, because the reproducibility between biological replicates is so high. Rather, this might reflect factors such as differences in protein and mRNA degradation, related to the adaptation of the two accessions to different environments.

Long-read genome assemblies and annotations are accurate but not yet perfect

Comparisons of our assemblies of Col-0 and Can-0 with published long-read assemblies of these accessions show that although they agree with high fidelity over most of the genome, some differences, concentrated within tandem repetitive regions remain. Typically, assemblies of the reference accession Col-0 disagree at about a few thousand SNP positions (Fig. 3). These discrepancies are partially due to algorithmic differences in the assembly pipelines but might also reflect some genuine variation in the germplasm sequenced in each study. Similarly, the identification of gene orthologs

between accessions depends on how paralogy within an accession is defined; orthologous genes can have different numbers of isoforms and so are arguably no longer true orthologs.

The Central Dogma revisited

Our study is incompatible with a naive interpretation of the Central Dogma's flow of information; it is not true that all the relevant information in CDS is mediated via mRNA abundance levels alone to modulate protein levels. One potential explanation is that our measurements of mRNA and protein expression are at single time points but represent the difference between synthesis and degradation integrated over the recent past. Potentially more information would be available if expression were measured at different time points due to, e.g., diurnal cycles. Some alternative explanations seem unlikely: First, our ribo-Seq levels are no better correlated with protein abundance than are standard RNA-seq levels, providing no support for the hypothesis that we are somehow biased away from relevant mRNA expression measurements by including transcripts not associated with ribosomes. Second, our data are from bulk tissue and not single cells. If it were technically possible to measure mRNA and protein from the same cells, we might observe stronger correlation between mRNA and protein, but it is difficult to see how this would explain how the additional information in CDS is actioned, because the latter is constant across cells. Third, tRNA abundance cannot explain the discrepancy.

Our CDS modeling summarizes just the gene codon frequencies, implicitly modeling gene length as well as gene amino acid frequencies and codon usage bias. As Fig. 9 shows, all three of these components are informative. We must therefore conclude that these codon frequencies are relevant to protein expression but in a manner not mediated by transcript levels alone.

It is known that on average protein and mRNA expression reduces with gene length [32], and our data supports this, but like other studies we report mRNA and protein abundance levels as normalized estimates of the numbers of copies of these molecules, independent of their length. This might explain why the component of CDS due to gene length is so relevant in our models. We do not have a complete mechanistic explanation for why amino acid frequencies and codon usage are also important. Regardless, since all the codon information in a CDS is also present in the sequence of its mRNA transcript, the Central Dogma still holds if the definition of information includes both abundance level and sequence. In a sense, this is closer to Francis Crick's original conception [49].

Conclusions

This study demonstrates how simple sequence features underlie much of the variation in expression of different omic levels and, in part, how these levels depend on each other, when and where it is helpful to measure a level, and when we can substitute one that is difficult to measure by a more tractable alternative. In particular, we have shown that, in the absence of environmental stress, the levels of constitutive methylation and tRNA abundance and their effects on expression are consequences of underlying sequence features. Finally, the Central Dogma's flow of information must be treated as multifactorial.

Methods

Plant growth

Seeds of the *Arabidopsis thaliana* accessions Col-0 and Can-0 were obtained from the Eurasian Stock Centre (formerly NASC). Seeds were sown on pre-soaked Levington F2 plus sand mix and stratified for 5 days at 5 °C before transferring to a controlled environment room. For long-read whole genome sequencing, plants were grown in short-day conditions: 10 h under fluorescent bulbs at 250 μmol/m^2/s, 23 °C and 65% relative humidity, followed by 14 h darkness at 18 °C with 75% relative humidity, and harvested shortly before bolting. Unless otherwise indicated, for all other assays, plants were grown in long-day conditions: 16 h under LED Lights at 150 μmol/m^2/s, 22 °C and 65% relative humidity, followed by 8 h darkness at 18 °C with 75% relative humidity and harvested at the 9-leaf stage.

Illumina whole genome DNA sequencing

DNA from a single leaf of each accession was extracted using the DNeasy Plant Mini kit (Qiagen, Manchester, UK) with on-column RNase treatment. Library preparation and Illumina sequencing (150 bp paired end reads) was performed by the Earlham Institute, UK. Read quality was checked using *Fastqc* [50], and sequencing adapters were trimmed using *BBduk* [51] with options "*ktrim = r k = 23 mink = 11 hdist = 1 tpe tbo*." To retain only good quality reads, a further round of quality trimming was conducted by *BBduk* with options "*qtrim = r trimq = 10 minlen = 50*."

Long-read whole genome DNA sequence

Col-0 and Can-0 plants were grown under short-day conditions until shortly before bolting and a single rosette for each genotype (~2 g tissue) was harvested and snap frozen in liquid nitrogen. High-molecular weight (HMW) DNA was extracted using a NucleoBond HMW DNA extraction kit (Macherey–Nagel, Dueren, Germany) as per the manufacturer's instructions. DNA quality control was performed using agarose gel electrophoresis, UV spectrophotometry (NanoDrop Technologies, USA), and the FP-1002 Genomic DNA 165 kb kit for FEMTO Pulse systems (Agilent technologies, Stockport, UK). DNA was quantified using Qubit high sensitivity DNA quantification kit (Thermo Fisher, Altrincham, UK). HMW DNA was sent to the Long Read Sequencing Facility (LRS) at University College London (London, UK) for library preparation and sequencing.

Libraries for sequencing using Oxford Nanopore Technologies (Oxford, UK, hereafter abbreviated to ONT or Nanopore) were prepared by using the ONT SQK-LSK109 ligation kit and sequenced on an ONT PromethION instrument. Initial sequencing of Can-0 produced an excessive number of short reads (< 10 kbp); therefore, the Short Read Eliminator Kit (SS-100–101-01; Circulomics Inc, USA) was used to progressively remove reads < 25 kbp in the Col-0 library before sequencing. Basecalling for ONT data was conducted using *guppy_basecaller* function in *Guppy* 5.0.16 [52], with options "*-c dna_r9.4.1_450bps_sup_prom.cfg –min_qscore 9 –min_score 40*

–trim_barcodes." We obtained 22.5 Gb data for Can-0 (150x), with a read N50 value around 26 kbp, and 30.3 Gb data (202x) for Col-0, with read N50 value around 33 kbp.

LRS also prepared libraries for PacBio HiFi sequencing (Pacific Biosciences of California, Inc. hereafter PacBio HiFi). DNA fragments were firstly sheared to ~17 kb by using Megaruptor 3 (Diagenode Inc.), and then the SMRTbell® Express Template Prep Kit 2.0 was used for constructing sequencing libraries. Libraries were sequenced on a PacBio Sequel II. Basecalling was performed by *SMRT Link* version 9.0.0.92188 [53] followed by circular consensus sequencing (CCS) analyses to generate HiFi sequences. For Can-0, we obtained 2.9 Gbp HiFi Q20 reads with a read N50 value of 14.5 kbp, and for Col-0 1.45 Gbp HiFi Q20 reads with a read N50 value of 14 kbp.

Genome assembly

For each accession, we first assembled the ONT and HiFi long-read data separately and then merged the assemblies. Nanopore data were initially assembled by using NECAT [54] with default settings except that "*CNS_OUTPUT_COVERAGE*" was changed to 40. The contigs generated by NECAT were polished by Nanopore reads using *Racon v1.4.20* [55] with options "*-m 8 -x −6 -g −8 -w 500*," followed by *Medaka 1.4.4* [56] with options "*-m r941_prom_sup_g507*," and finally polished with Illumina reads by *Pilon* 1.24 [57] with options "*–changes –fix all*." In all the polishing steps, *minimap2* [58] was used to map long reads to the contigs, and *bwa-mem2* [59] was used to map Illumina short-read sequences. PacBio HiFi reads were assembled by *hifiasm* with option "-l 0."

To combine the higher continuity from Nanopore-based contigs with the higher accuracy from PacBio HiFi-based contigs, we used *quickmerge* [60] setting the Nanopore assemblies (after three polishing steps described above) as "*hybrid_assembly*" and the HiFi-read assemblies as "*self_assembly*," and the N50 value of the polished Nanopore-read assemblies as the minimum length cutoff (-l) of contigs to be merged. After the merging, we performed two rounds of polishing the merged assemblies with PacBio HiFi reads using racon. The alignment of HiFi reads to the merged assemblies was conducted with *pbmm2* [61]. Where the merged assembly did not retain a similar N50 value to the Nanopore-read assembly, we patched the Nanopore assembly to the merged assembly using *RagTag patch* [62]. Finally, polished merged assemblies were placed on the correct chromosomes using the published Col-0-CEN assembly [17] by using *RagTag scaffold* [62]. After scaffolding, we ran another round of final polishing of the scaffold by haplotype-aware polishing tool *Hapo-G* [63]. At each round of polishing and merging, the N50 values of assemblies or scaffolds were estimated by *QUAST* [64], the quality value (QV) of assemblies or scaffolds estimated by *Merqury* [65], and the number of Benchmarking Universal Single-Copy Orthologs (BUSCO) for estimating completeness and duplication calculated by *BUSCO* [66, 67]. Every round of polishing increased the QV of the assemblies or scaffolds.

Omni-C sequence analysis

Genome-wide chromatin interaction data for Col-0 and Can-0 was generated using a Dovetail® (now Cantata Bio, USA) Omni-C® Kit. Plants were grown under long-day conditions for 19 days and then dark-treated for 48 h before leaf tissue was snap frozen in liquid nitrogen. Five hundred milligrams of tissue was ground into a fine powder

with liquid nitrogen using a mortar and pestle and the assay performed as per the manufacturer's instructions for plants. Briefly, chromatin was fixed with formaldehyde and nuclease treated. An aliquot was de-crosslinked, and DNA purified with a DNA Clean & Concentrator-5 kit (Zymo Research, Irvine, CA, USA). DNA yield and fragment size were determined using a Bioanalyzer high sensitivity dsDNA kit (Agilent Technologies, Milton Keynes, UK) and Qubit high sensitivity DNA quantification kit (Thermo Fisher, Altrincham, UK). The remaining lysate was then processed with reactions for end-polishing, ligation of a biotinylated oligonucleotide bridge, intra-aggregate ligation, and cross-link reversal, respectively. The DNA was purified and quantified using a Qubit high sensitivity DNA quantification kit (Thermo Fisher, Altrincham, UK) before proceeding with library preparation. Streptavidin enrichment of the biotinylated bridge was performed, and the final libraries were indexed and amplified by PCR. Illumina NovaSeq PE150 sequencing was performed by Novogene Co. Ltd (Cambridge, UK), on a Novaseq 6000 instrument. The 150 bp paired-end Omni-C reads were aligned to the merged and polished scaffolds. Read deduplication and finding contact points was performed by following the Dovetail Omni-C kit document at https://omni-c.readthedocs.io/en/latest/index.html. *PretextMap* [68] and *PreTextView* [69] were used to generate and view the final contact map, to monitor and confirm the structure of scaffolds. Supplemental Figures S1 and S2 were generated from the contact maps using *PretextSnapshot* [70].

ONT methylation data generation and differential methylation analysis

Because of the larger amount of leaf tissue required, plants for ONT sequencing were grown under short-day conditions, whereas long-day conditions were used for bisulfite sequencing to enable direct comparisons with the RNA-seq and proteomics datasets. ONT fast5 files were processed with *Megalodon v2.4.1* [71] and *Guppy* (GPU version *5.0.16_linux64*) [52] with the option "*–guppy-config dna_r9.4.1_450bps_sup_prom.cfg –remora-modified-bases dna_r9.4.1_e8 Sup 0.0.0 5mc CG 0*" to generate the raw methylation data. The R package *NanoMethViz* [72] was used to visualize the ONT methylation data, and as well to prepare input files for *DSS* [73] for differential methylation calling between Col-0 and Can-0. The threshold p value for calling differentially methylated loci (i.e., at individual nucleotides) was set to 0.01, and the threshold of p value for calling differentially methylated regions (e.g., across gene bodies) was 0.05. The correlation between bisulfite and nanopore methylation results was performed by the "*megalodon_extras validate compare_modified_bases*" function in the *megalodon* package. We quantified each methylated CpG dinucleotide as the percentage of methylated bases from reads covering that CpG position.

Comparative multi-omic analyses

Comparative assays were performed on Can-0 and Col-0 grown under long-day conditions (described above) until the emergence of the 9th rosette leaf. Leaves 3–8 were harvested and frozen in liquid nitrogen. Each biological replicate comprised five plants, pooled and homogenized by grinding with a mortar and pestle in liquid nitrogen.

Whole-genome bisulfite sequencing (WGBS)

DNA was extracted from three biological replicates of pooled leaf tissue using a DNeasy Plant Mini Kit (Qiagen, Manchester, UK). WGBS Libraries were constructed from 10 ng DNA using a Pico Methyl-Seq™ Library Prep Kit for Illumina-based Sequencing (Zymo Research, Irvine, CA, USA). DNA quality control was performed using agarose gel electrophoresis and UV spectrophotometry (NanoDrop Technologies, USA). Quantification of extracted DNA was performed using a Qubit high sensitivity DNA quantification kit (Thermo Fisher, Altrincham, UK). The quality and quantity control of WGBS Libraries employed an Agilent DNA 1000 Kit (Agilent technologies, Milton Keynes, UK) and Agilent 2100 Bioanalyzer. All kit workflows were performed according to manufacturers' instructions. Libraries were sequenced by Novagene (150 bp paired end reads, using a NovaSeq instrument). The bisulfite reads were firstly trimmed by 10 bp at each end by *trim_galore* [74], and then *bismark* [75] was run for read mapping, deduplication, and extraction of methylation information. The script *dname_bed_corr.sh* from *dna_me_pipeline* [75] was used to check the correlation of methylation between samples. Loci covered by fewer than 10 reads were removed.

Short-read transcriptome sequencing (RNA-seq)

RNA was extracted from five biological replicates of pooled leaf tissue using a Plant RNeasy Mini Kit (Qiagen, Manchester, UK) as per the manufacturer's recommendations. DNase treatment was performed using a TURBO DNA-free kit (Invitrogen, now ThermoFisher Scientific). RNA quantification and analysis of integrity employed an Agilent RNA 6000 Nano Kit (Agilent technologies, Milton Keynes, UK) and the Agilent 2100 Bioanalyzer. Illumina NovaSeq PE150 sequencing was performed by Novogene Co. Ltd, UK. The sequencing data were trimmed and filtered by *BBduk* [51] with options "*qtrim=rl trimq=10 maq=10*" to achieve clean and good quality sequences. mRNA expression levels were summed across all isoforms for a given gene using *kallisto* [76] to produce normalized transcripts per million (TPM) values within each replicate; the expression of a gene across replicates was estimated by the $\log_{10}$ of their Geometric mean, after adding a pseudocount of 1.0 to avoid negative infinite values.

tRNA quantification (mim-tRNAseq)

RNA was isolated from two biological replicates of pooled leaf tissue using phenol/chloroform extraction. Tissue was cryo-pulverized in liquid nitrogen using a Geno/Grinder® (Spex SamplePrep 2010 USA) and 1 mL of TRIzol™ Reagent was added to 300–400 µL tissue powder, followed by incubation at RT for 5 min. RNA was extracted by the addition of 0.2 vol chloroform, incubation at RT for 2 min and centrifugation at 12,000 g for 15 min at 4 °C. The upper aqueous phase was extracted with an equal volume of chloroform and the RNA precipitated by addition of ice-cold 100% ethanol. Following centrifugation at 12,000 g for 20 min at 4 °C, the pellet was washed in 80% (v/v) ethanol, air-dried, and resuspended in RNAse-free water. RNAs were sequenced using modification-induced misincorporation tRNA sequencing (mim-tRNAseq) [34, 35].

Data analysis was performed using the bioinformatics pipeline in [34]. In brief, the sequences were trimmed with *cutadapt* version 4.1 [77]; a first step trims the GATATC

GTCAAGATCGGAAGAGC adapter in 3′, a second step trims the two bases due to the circularization (*-u 2*), and a last run cuts the remaining adapter in 5′: CTTGACGAT ATC). Only reads longer than 25 bp with quality > 25 were retained for mim-tRNAseq analysis. The package *mim-tRNAseq* version 1.1.7 [34] was used with the parameters "*–species Atha–cluster-id 0.95–threads 15–min-cov 0.0005–max-multi 4–remap–remap-mismatches 0.075*." We determined the Araport11 ids of the resulting tRNA genes based matching their TAIR10 coordinates.

Ribosome profiling

Ribonucleic complexes from each accession were solubilized from six biological replicates of pooled leaf tissue and clarified as per [25]. Polysomes were separated on a Sucrose gradient with absorbance measured at 254 nm using a UV-1 monitor (Pharmacia, Uppsala, Sweden). Ribosome profiling (3′Riboseq) was performed as described in [78] with pooling of monosome and polysomes. Ribosome-associated RNA was precipitated using sodium acetate and ethanol and purified using a Zymo quick RNA column (Zymo Research, Irvine, CA) and the integrity assessed using an Agilent Bioanalyzer 2100. mRNA Libraries were prepared by Novogene Co. Ltd., UK, and sequenced using NovaSeq to produce paired end 150 bp reads. Data was analyzed as for RNA-seq, with additional filtering to remove ribosomal RNAs and tRNAs using sequences from [79].

Quantitative proteomics

Protein extraction, reduction, alkylation, and digestion

Total protein was extracted from four biological replicates of pooled leaf tissue. Tissue samples (100 mg) were cryo-pulverized in Liquid nitrogen using a mortar and pestle. Protein was precipitated by the addition of 5 mL pre-chilled 10% (w/v) trichloroacetic acid (TCA) in acetone, followed by incubation at −20 °C overnight. The precipitated protein pellet was washed three times with chilled acetone. TCA-precipitated pellets were solubilized and reduced in 100 µL of urea containing buffer [8 M urea, 50 mM triethylammonium bicarbonate (TEAB), 10 mM dithiothreitol, 1× protease inhibitor cocktail (Roche, Mannheim, Germany)] at 25 °C for 1 h. Protein was alkylated by the addition of 2-chloroacetamide to achieve a final concentration of 55 mM, followed by incubation at 25 °C for 30 min. In-solution protein digestion was performed in three sequential steps: for the first digest, 2 µg of Lys-C (Promega, Madison, WI) was added to give a 1:100, w/w enzyme-protein ratio, followed by incubation at 37 °C for 4 h. For the second digest, 700 µL of 50 mM TEAB was added to reduce the urea concentration below 1 M, and 2 µg of trypsin (Promega) was added. The mixture was incubated at 37 °C overnight, followed by the addition of a further 2 µg of trypsin for the third digest and a 4-h incubation at 37 °C. The digests were acidified with 1% trifluoroacetic acid (TFA) and desalted using 50 mg SepPak tC18 cartridges (Waters Corporation, Borehamwood, UK). The cartridge was washed with 0.1% TFA solution and eluted in two steps: (i) 300 µL 25% acetonitrile (ACN) in 0.1% formic acid (FA); (ii) twice with 300 µL 50% ACN in 0.1% FA. The eluates were lyophilized and stored at −20 °C. Peptide amount was determined using a Pierce Quantitative Colorimetric Peptide Assay kit (Thermo Fisher Scientific, Hemel Hempstead, UK).

Mass spectrometry

The proteomics data was acquired using a timsTOF HT mass spectrometer (Bruker Daltonics, Bremen, Germany) coupled with a nanoElute 2 UHPLC system (Bruker Daltonics). Peptides (750 ng) were loaded onto a PepMap Neo trap column (300 µm × 5 mm, 5 µm particle size, Thermo Scientific) and separated on a µPAC Neo analytical column (500 mm × 180 µm, 16 µm pillar length, Thermo Scientific) using a 60-min non-linear gradient consisting of 5%–17% solvent B over 42 min at a flow rate of 300 nL/min, followed by an increase to 26% for 14 min and 37% for 4 min. The mobile phases comprised 0.1% FA in water as solvent A and 0.1% FA in ACN as solvent B. The eluates were ionized using a Captive Spray source via a ZDV Sprayer emitter (20 µm, Bruker Daltonics). The mass spectrometer was set to dia-PASEF scan mode spanning 100–1700 m/z in positive ion mode. The ion mobility (IM) range was set to 0.85–1.23 1/K0 [V s/cm^2], and both the ramp time and the accumulation time were set to 100 ms, corresponding to a ramp rate of 9.42 Hz. The variable collision energy was applied depending on the IM, ranging from 20 eV at 0.60 1/K0 to 59 eV at 1.6 1/K0. The dia-PASEF windows were optimized for the *Arabidopsis* proteome profile using *py_diAID* version 0.0.19 [80]. Ten dia-PASEF scans were divided into 3 IM windows with a mass range of 300–1200 Da, corresponding to an estimated cycle time of 1.17 s.

Proteome data analysis

The mass spectra from Col-0 and Can-0 were searched separately against their own sequence databases using *DIA-NN v1.8.2 beta 27* [81]. The Col-0 and Can-0 protein sequence databases derived from de novo assemblies were used to generate in silico spectral libraries, which contain predicted retention times and predicted ion mobility (1/K0) values. A maximum of one missed cleavage was permitted with a minimum peptide length of 7 amino acids. Dynamic modifications considered oxidized methionine, acetylation at the protein N-terminus, and methionine loss at the protein N-terminus. Carbamidomethylation of cysteine was designated as a fixed modification. The "*match between runs*" option was utilized to minimize missing identifications. The R package *DIAgui* version *1.4.2* [82] was used to generate iBAQ values [22] from the protein intensity, filtered at both precursor and gene levels at 1% FDR, using only proteotypic peptides. The protein matrices from Col-0 and Can-0 were combined using the Hierarchical Orthologous Group (HOG) classification, which identifies orthologous genes between accessions (see below). The iBAQ value for gene i was normalized using the equation:

$$\text{normalized iBAQ}(i) = \frac{\text{iBAQ}(i)}{\sum \text{iBAQ}} \times 10^9$$

Long-read transcriptome sequencing (Iso-seq)

Iso-seq data were generated from Can-0 and Col-0 rosette leaves grown under long-day conditions (described above) to support genome annotation only. Total RNA was extracted from one biological replicate of pooled leaf tissue, using a Plant RNeasy

Mini Kit (Qiagen, Manchester, UK) as per the manufacturer's recommendations. Quantification of RNA and analysis of integrity was performed using an Agilent RNA 6000 Nano Kit (Agilent Technologies, Stockport UK). PacBio Iso-Seq™ SMRTbell® Libraries were constructed by Novogene Co. Ltd, UK, with the Express Template Prep Kit 2.0 for sequencing on Sequel® II System. Sequencing data were analyzed according to the *Isoseq3* instructions [83], including consensus sequence generation by *ccs* [84], primer removal and demultiplexing by *lima*, trimming the polyA tail and concatemer by *isoseq*. The clustering step suggested by the Isoseq3 workflow was omitted.

Genome annotation

The finalized genome assemblies of Col-0 and Can-0 were first scanned by *EDTA* [85] with options "*–sensitive 1 – evaluate 1 –anno 1*" to mask and annotate repeats, using the *RepBase* [86] transposable element (TE) database (version 27.01) of *A. thaliana* as the curated TE library and the known coding sequences of *A. thaliana* (Araport11 [20]). The sequences of annotated repeats were aligned against the Araport11 coding sequences using *blastn* [87], to ensure that no gene sequences overlapped with the repeats identified by *EDTA*.

After cleaning and QC, the Illumina RNAseq reads for Col-0 and Can-0 were each aligned to their respective assembled genomes by *STAR* [88] with option "*–outSAMstrandField intronMotif*" to generate RNAseq alignment.*bam* files. The.*bam* files together with the soft-masked genomes generated by *EDTA* were then input to *Braker1* [89] for gene prediction and annotation. We used the option "*–UTR=on –augustus_args="–species=arabidopsis""* to apply the *Augustus* [90] pre-trained gene model of *Arabidopsis* for ab initio gene prediction and UTR annotation. We used *Braker2* to annotate genes using the Uniprot *A. thaliana* proteome database [91], and the long-read protocol from *Braker* to generate a version of annotation that incorporated our PacBio Isoseq data. We mapped the filtered Iso-Seq reads to the genome by *minimap2* [92] with option "*-ax splice -uf –secondary=no -C5*" following the long-read protocol from *Braker* [93]. The three gene annotations respectively from *Braker1*, *Braker2*, and long-read protocol were then merged and filtered by *Tsebra* [94] with the option "*long_reads_filtered.cfg*."

We used the *PASA* pipeline [95] to refine these *Tsebra* annotations. First, *Trinity* [96] produced genome-guided and de novo assembled transcriptomes from both Isoseq and Illumina RNAseq. Secondly, we ran the *PASA* pipeline using default options with two rounds of annotation updates on the Tsebra annotation. To make sure that we have the most complete gene set for the accessions, we also augmented our updated annotations with the lifted over annotation from Tair10 which was produced by *Liftoff* [97] with default settings. This final step was performed by the function *agat_sp_complement_annotations.pl* from *AGAT* v.0.9.2 [98] using the *PASA* updated annotation as the reference and adding any additional genes from lifted over annotations. Functional annotation for the finalized annotations was conducted by *InterProScan* [99] on https://usegalaxy.eu [100]. We used *AGAT* to integrate functional and homology annotations into GFF format. Finally, we used *BUSCO* to estimate the completeness of the annotated transcriptomes, and *AGAT* to correct small problems in the annotation such as duplication of genes with different identifiers and changing gene IDs.

We identified orthologous genes using *OMA standalone* [101]. We downloaded the OMA database orthologs from the five Brassica species (*Arabis alpina*, *Arabidopsis lyrate*, *Arabidopsis thaliana (Tair10)*, *Brassica napus*, *Brassica oleracea*, *Brassica rapa* subsp. *Pekinensis*) and combined them with the TAIR10, Col-0, and Can-0 protein sequences into Hierarchical Orthologous Groups (HOG) using *pyHam* [102]. For HOGs containing more than one gene from the same accession, gene DNA sequences were used to identify the most similar homologs inter-accessions by reciprocal *blastn* [87, 103] searches. If were more than two genes from each same accession in a HOG, we used DNA sequence of each gene to search for the genes pairs between accessions that are the most similar, to find the 1-to-1 homologous genes.

Quality control of gene annotations

We applied quality control filters to remove duplicated genes, genes longer than 6 kb, and genes without a 1–1 ortholog between the two accessions, leaving 22,606 genes. Of these genes, 18,226 were expressed in Col-0 (and 772 only in Col-0) and 18,200 expressed in Can-0 (746 only in Can-0). Here a gene is "expressed" if at least one RNAseq read aligned unambiguously to its gene sequence in at least one biological replicate. There were 17,454 ortholog pairs with mRNA expression in at least one replicate in both accessions. After removing genes with premature stop codons, 17,414 expressed genes remained for analysis. If we further only retain those genes with expressed mRNA in all five biological replicates in each accession, then there are 15,669 in Col-0, 15,669 in Can-0, and 15,215 in both. Our downstream analyses do not employ this additional filter except where noted.

Comparative genomics

Our assemblies of Col-0 and Can-0 were aligned using *minimap2* [92]. Synteny statistics were calculated *by dna_diff* [19]. High-confidence variants including indels below 10 bp were called by *clair3* [104] and structural variants were called using *pbsv* [105] with HiFi reads.

Definition of expressed genes

For mRNA, we define a gene to be expressed if it is detected by alignment of Our mRNAseq data to our gene models. Since Our RNAseq data have sequencing saturation rates of 63% to 71%, we believe that only a few expressed genes expressed are not detected.

For protein, expressed means proteins quantified by MS. Low-abundance proteins under the limit of detection or those with physiochemically undetectable sequences were not classified as expressed. Note that we do not apply any numerical cutoff to these definitions. However, we additionally report genes expressed in all biological replicates for certain analyses (e.g., Additional file 1: Fig. S3).

Differential expression analysis

Clean RNAseq data from Col-0 and Can-0 were fed into *Trinity v2.14.0* [96] for RNAseq quantification and differential expression analysis, using default parameters with Col-0 as the reference genome. We used *kallisto* [76] for pseudo-alignment of

RNAseq reads and quantification of gene and isoform expression. The output from *kallisto* was used as input for *EdgeR* [106] for normalization and differential expression analysis. The gene expression within each replicate sample was normalized by calculating transcripts per million (TPM), and inter-sample normalization was performed by calculating Trimmed Mean of M-values (TMM) in *EdgeR*. We used a FDR threshold of 0.05 to identify genes or transcripts that are differentially expressed between Col-0 and Can-0.

The normalized iBAQ proteome values were log-transformed and missing values imputed with random values from a normal distribution (width 0.3, down shift 1.8) using *Perseus v1.6.15.0* [107]. DE proteins were determined using Student *t*-test with threshold of 0.05 for the Benjamini–Hochberg adjusted *p* value (FDR) and a twofold change.

Testing relationships between annotations and expression

Relationships between a dichotomous annotation difference between Col-0 and Can-0 (such as differential gbM or the presence/absence of a TE upstream of a gene) and continuous mRNA or protein expression were tested by first subdividing the genes into two subsets corresponding to those genes with and without the specified attribute (e.g., whether or not the gene is differentially methylated) and then testing if the mean of the absolute values of the $\log_2$ fold change of mRNA or protein expression in the two subsets differed, using a non-parametric Mann–Whitney test implemented in R. Statistical significance was reported as logP, the negative $\log_{10}(p \text{ value})$ of the test. This methodology does not assume any particular direction of effect between annotation and expression.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03741-0.

Additional file 1: All supplemental figures and legends.

Additional file 2: Supplemental Tables S1 and S2 with legends.

Additional file 3: Table S3 Tabular text file listing orthologous genes between Col-0, Can-0, and Araport11. Each row represents one homologous group (HOG). Numbers of alternatively spliced isoforms annotated for each gene are indicated by the "Transcripts" columns.

Additional file 4: Table S4 Normalized expression values for gbM, mRNA, protein, and ribo-RNA. Values are supplied for each replicate (except for gbM) and combined across replicates, and log-transformed in both accessions Col-0 and Can-0.

Additional file 5: Table S5 Pearson correlations between replicates for log-transformed mRNA, protein, and ribo-RNA expression values. Light blue background indicates correlations between replicates within an accession; white background indicates correlations between replicates from different accessions.

Additional file 6: Table S6 Estimated codon effects in Col-0 and Can-0 together with the tRNA codon effects.

Additional file 7: Table S7 tRNA expression values for Col-0 and Can-0.

Additional file 8: Table S8 ANOVA tables used to generate Fig. 9.

Additional file 9: Table S9 Differential expression analysis.

Additional file 10: Table S10 Accession numbers of sequencing data submitted to the European Nucleotide Archive.

Additional file 11: Text T1 Analysis linking estimated codon effects to mRNA half-lives.

Acknowledgements

We thank Hongtao Zhang for preliminary data to support the funding application, Katrin Strasser for help with tRNA sequencing library preparation and Rob King for preliminary RNA-seq analysis. We also thank the NGS Facility in the Department of Totipotency at the MPI of Biochemistry and the Earlham Institute and the UCL Long Read Sequencing Facility for DNA and RNA sequencing and are grateful to Ian Henderson and Oxford Nanopore Technology technical support staff for advice on long read sequencing and genome assembly.

Peer review information
Wenjing She was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions
FLT, KSL, RM, KHP, GS, DN conceived the project and designed the experiments. MB, YK, NPA, BP, CR, DN performed the experiments. ZZ, RM, YK, LA, KHP, DB performed the analyses. RM, FLT, KSL, ZZ, YK, XL wrote or edited the paper.

Funding
This work was funded by BBSRC grants BB/T002182/1 (awarded to FLT, RM, and KSL), BB/X017877/1 (awarded to FLT), BB/W019620/1 (awarded to KSL and FLT), and by BB/W510543/1—21ROMITIGATIONFUND Rothamsted (to MB). DDN acknowledges funding by the Max Planck Society and the European Research Council under the European Union's Horizon 2020 Research and Innovation Programme (ERC Starting Grant number 803825-TransTempoFold).

Data availability
All Col-0 and Can-0 DNA and RNA sequence data and the genome assemblies are publicly available from ENA under the study accession number ERP161680 [21]. All samples and their accession numbers are listed in Supplemental Table S10. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [108] partner repository with the dataset identifier PXD058342 [109].
Annotation files:
Col-0 and Can-0 genome annotations, mRNA and peptide sequences are available from the public UCL Figshare repository with https://doi.org/10.5522/04/28163222 [21].
R code that generates the figures and tables:
An R markdown document and associated datasets are available from Zenodo under https://doi.org/10.5281/zenodo.15467917 [110]. This repository, when downloaded, extracted and executed (knitted) in RStudio will reproducibly create all the main figures except Fig. 1 (composite) and Fig. 12 (manually created), and perform the correlation and modelling analyses presented in the paper.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
Danny Nedialkova and Andrew Behrens are listed as inventors on a patent application filed by the Max Planck Society pertaining to the mim-tRNAseq technology. The authors confirm that a patent would not prevent the reuse of any material such as data or code, which is freely available on GitHub [35] under a GPL-3.0 license.

Author details
[1]Genetics Institute, University College London, London WC1E 6BT, UK. [2]Plant Sciences and the Bioeconomy, Rothamsted Research, Harpenden AL5 2JQ, UK. [3]Cambridge Centre for Proteomics, Department of Biochemistry, University of Cambridge, Cambridge, UK. [4]Leibniz Centre for Agricultural Landscape Research (ZALF), Müncheberg 15374, Germany. [5]Mechanisms of Protein Biogenesis, Max Planck Institute of Biochemistry, Martinsried 82152, Germany. [6]Department of Bioscience, TUM School of Natural Sciences, Technical University of Munich, Garching 85748, Germany. [7]Institute of Biological Environmental and Rural Sciences (IBERS), Aberystwyth University, Gogerddan, Aberystwyth, UK.

Received: 8 January 2025 Accepted: 12 August 2025
Published online: 29 September 2025

References

1. Ponnala L, Wang Y, Sun Q, van Wijk KJ. Correlation of mRNA and protein abundance in the developing maize leaf. Plant J. 2014;78:424–40.
2. Becker K, Bluhm A, Casas-Vila N, Dinges N, Dejung M, Sayols S, et al. Quantifying post-transcriptional regulation in the development of *Drosophila melanogaster*. Nat Commun. 2018;9:4970.
3. Maier T, Guell M, Serrano L. Correlation of mRNA and protein in complex biological samples. FEBS Lett. 2009;583:3966–73.
4. Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat Rev Genet. 2012;13:227–32.
5. Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL. Evolution of proteins and gene expression levels are coupled in *Drosophila* and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. Mol Biol Evol. 2005;22:1345–54.
6. Buccitelli C, Selbach M. mRNAs, proteins and the emerging principles of gene expression control. Nat Rev Genet. 2020;21:630–44.
7. Outeiral C, Deane C. Codon language embeddings provide strong signals for use in protein engineering. Nat Mach Intell. 2024;6:170–9.
8. Takuno S, Gaut BS. Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. Proc Natl Acad Sci U S A. 2013;110:1797–802.

9. Muyle A, Ross-Ibarra J, Seymour DK, Gaut BS. Gene body methylation is under selection in *Arabidopsis thaliana*. Genetics. 2021;218:iyab061.
10. Schmid MW, Heichinger C, Coman Schmid D, Guthorl D, Gagliardini V, Bruggmann R, et al. Contribution of epigenetic variation to adaptation in Arabidopsis. Nat Commun. 2018;9:4446.
11. Muyle AM, Seymour DK, Lv Y, Huettel B, Gaut BS. Gene body methylation in plants: mechanisms, functions, and important implications for understanding evolutionary processes. Genome Biol Evol. 2022. https://doi.org/10.1093/gbe/evac038.
12. Zilberman D. An evolutionary case for functional gene body methylation in plants and animals. Genome Biol. 2017;18:87.
13. Somssich M. A short history of Arabidopsis thaliana (L.) Heynh. Columbia-0. PeerJ Preprints. 2019;7:e26931v26935.
14. Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, et al. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. Nature. 2011;477:419–23.
15. Consortium G. 1,135 genomes reveal the global pattern of polymorphism in Arabidopsis thaliana. Cell. 2016;166:481–91.
16. Lian Q, Huettel B, Walkemeier B, Mayjonade B, Lopez-Roques C, Gil L, et al. A pan-genome of 69 *Arabidopsis thaliana* accessions reveals a conserved genome structure throughout the global species range. Nat Genet. 2024;56:982–91.
17. Naish M, Alonge M, Wlodzimierz P, Tock AJ, Abramson BW, Schmucker A, et al. The genetic and epigenetic landscape of the Arabidopsis centromeres. Science. 2021;374:eabi7489.
18. Wang B, Yang X, Jia Y, Xu Y, Jia P, Dang N, et al. High-quality *Arabidopsis thaliana* genome assembly with Nanopore and HiFi long reads. Genomics Proteomics Bioinformatics. 2022;20:4–13.
19. Delcher AL, Salzberg SL, Phillippy AM. Using MUMmer to identify similar regions in large sequence sets. Curr Protoc Bioinformatics. 2003; Chapter 10:Unit 10 13.
20. Cheng CY, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD. Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. Plant J. 2017;89:789–804.
21. Zhong Z, Bailey M, Kim Y, Afsharyna NP, Parker B, Arathoon L, Li X, Rundle CA, Behrens A, Nedialkova D, Slavov G, Hassani-Pak K, Lilliey KS, Theodoulou FL, Mott R. Revisiting the Central Dogma: the distinct roles of genome, methylation, transcription, and translation on protein expression in Arabidopsis thaliana. UCL Figshare Repository 2025.
22. Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, et al. Global quantification of mammalian gene expression control. Nature. 2013;495:126–7.
23. Zubarev RA. The challenge of the proteome dynamic range and its implications for in-depth proteomics. Proteomics. 2013;13:723–6.
24. Munro V, Kelly V, Messner CB, Kustatscher G. Cellular control of protein levels: a systems biology perspective. Proteomics. 2024;24:e2200220.
25. Merchante C, Hu Q, Heber S, Alonso J, Stepanova AN. A ribosome footprinting protocol for plants. Bio-Protoc. 2016;6:e1985.
26. Guanzon D, Ross JP, Ma C, Berry O, Liew YJ. Comparing methylation levels assayed in GC-rich regions with current and emerging methods. BMC Genomics. 2024;25:741.
27. Zhou Z, Dang Y, Zhou M, Li L, Yu CH, Fu J, et al. Codon usage is an important determinant of gene expression levels largely through its effects on transcription. Proc Natl Acad Sci U S A. 2016;113:E6117–25.
28. Quax TE, Claassens NJ, Soll D, van der Oost J. Codon Bias as a Means to Fine-Tune Gene Expression. Mol Cell. 2015;59:149–61.
29. Liu Y, Yang Q, Zhao F. Synonymous but not silent: the codon usage code for gene expression and protein folding. Annu Rev Biochem. 2021;90:375–401.
30. Yang Q, Lyu X, Zhao F, Liu Y. Effects of codon usage on gene expression are promoter context dependent. Nucleic Acids Res. 2021;49:818–31.
31. Ehrlich R, Davyt M, Lopez I, Chalar C, Marin M. On the track of the missing tRNA genes: a source of non-canonical functions? Front Mol Biosci. 2021;8:643701.
32. Berg MD, Brandl CJ. Transfer RNAs: diversity in form and function. RNA Biol. 2021;18:316–39.
33. Gao L, Behrens A, Rodschinka G, Forcelloni S, Wani S, Strasser K, et al. Selective gene expression maintains human tRNA anticodon pools during differentiation. Nat Cell Biol. 2024;26:100–12.
34. Behrens A, Rodschinka G, Nedialkova DD. High-resolution quantitative profiling of tRNA abundance and modification status in eukaryotes by mim-tRNAseq. Mol Cell. 2021;81(1802–1815):e1807.
35. Behrens A, Nedialkova DD. Experimental and computational workflow for the analysis of tRNA pools from eukaryotic cells by mim-tRNAseq. STAR Protoc. 2022;3:101579.
36. Chan PP, Lowe TM. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. Nucleic Acids Res. 2016;44:D184-189.
37. Rogers DW, Bottcher MA, Traulsen A, Greig D. Ribosome reinitiation can explain length-dependent translation of messenger RNA. PLoS Comput Biol. 2017;13:e1005592.
38. Ge SX, Jung D, Yao R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. Bioinformatics. 2020;36:2628–9.
39. Barbadilla-Martinez L, Klaassen N, van Steensel B, de Ridder J. Predicting gene expression from DNA sequence using deep learning models. Nat Rev Genet. 2025.
40. Benegas G, Ye C, Albors C, Li JC, Song YS. Genomic language models: opportunities and challenges. Trends Genet. 2025;41:286–302.
41. Hanson G, Coller J. Codon optimality, bias and usage in translation and mrna decay. Nat Rev Mol Cell Biol. 2018;19:20–30.
42. Presnyak V, Alhusaini N, Chen YH, Martin S, Morris N, Kline N, et al. Codon optimality is a major determinant of mRNA stability. Cell. 2015;160:1111–24.
43. Radhakrishnan A, Green R. Connections underlying translation and mRNA stability. J Mol Biol. 2016;428:3558–64.

44. Vogel C, Abreu Rde S, Ko D, Le SY, Shapiro BA, Burns SC, et al. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. Mol Syst Biol. 2010;6:400.
45. Agarwal V, Kelley DR. The genetic and biochemical determinants of mRNA degradation rates in mammals. Genome Biol. 2022;23:245.
46. Chu D, Wei L. Characterizing the heat response of *Arabidopsis thaliana* from the perspective of codon usage bias and translational regulation. J Plant Physiol. 2019;240:153012.
47. Kover PX, Valdar W, Trakalo J, Scarcelli N, Ehrenreich IM, Purugganan MD, et al. A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana*. PLoS Genet. 2009;5:e1000551.
48. Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. Science. 2009;324:218–23.
49. Crick F. Central dogma of molecular biology. Nature. 1970;227:561–3.
50. Fastqc, a quality control tool for high throughput sequence data. [https://www.bioinformatics.babraham.ac.uk/projects/fastqc/].
51. BBMap short read aligner, and other bioinformatic tools. [https://sourceforge.net/projects/bbmap/].
52. Guppy. [https://community.nanoporetech.com/docs/prepare/library_prep_protocols/Guppy-protocol/v/gpb_2003_v1_revax_14dec2018/guppy-software-overview].
53. Inc PB: SMRT Link. 2023. https://www.pacb.com/smrt-link/.
54. Chen Y, Nie F, Xie SQ, Zheng YF, Dai Q, Bray T, et al. Efficient assembly of nanopore reads via highly accurate and intact error correction. Nat Commun. 2021;12:60.
55. Vaser R, Sovic I, Nagarajan N, Sikic M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 2017;27:737–46.
56. Technologies ON. Medaka. 2023. https://github.com/nanoporetech/medaka.
57. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE. 2014;9:e112963.
58. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.
59. Md V, Misra S, Li H, Aluru S. Efficient architecture-aware acceleration of BWA-MEM for multicore systems. In: IEEE Parallel and Distributed Processing Symposium; Rio de Janeiro, Brazil. IEEE. 2019:314–324.
60. Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. Nucleic Acids Res. 2016;44:e147.
61. Toepfer A. pbmm2: A minimap2 SMRT wrapper for PacBio data. 2023. https://github.com/PacificBiosciences/pbmm2.
62. Alonge M, Lebeigle L, Kirsche M, Jenike K, Ou S, Aganezov S, et al. Automated assembly scaffolding using ragtag elevates a new tomato system for high-throughput genome editing. Genome Biol. 2022;23:258.
63. Aury JM, Istace B. Hapo-G, haplotype-aware polishing of genome assemblies with accurate reads. NAR Genomics Bioinform. 2021;3:lqab034.
64. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29:1072–5.
65. Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 2020;21:245.
66. Manni M, Berkeley MR, Seppey M, Zdobnov EM. Busco: assessing genomic data quality and beyond. Curr Protoc. 2021;1:e323.
67. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2.
68. PretextMap. [https://github.com/wtsi-hpag/PretextMap].
69. PretextView. [https://github.com/wtsi-hpag/PretextView].
70. Harry E: PretextSnapshot. 2021. https://github.com/sanger-tol/PretextSnapshot.
71. Megalodon v2.4.1. [https://nanoporetech.github.io/megalodon/].
72. Su S, Gouil Q, Blewitt ME, Cook D, Hickey PF, Ritchie ME. NanoMethViz: an R/bioconductor package for visualizing long-read methylation data. PLoS Comput Biol. 2021;17:e1009524.
73. Feng H, Wu H. Differential methylation analysis for bisulfite sequencing using DSS. Quant Biol. 2019;7:327–34.
74. TrimGalore. [https://github.com/FelixKrueger/TrimGalore].
75. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011;27:1571–2.
76. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34:525–7.
77. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011;2011(17):3.
78. Zhu W, Xu J, Chen S, Chen J, Liang Y, Zhang C, et al. Large-scale translatome profiling annotates the functional genome and reveals the key role of genic 3′ untranslated regions in translatomic variation in plants. Plant Commun. 2021;2:100181.
79. Wu HL, Hsu PY. A custom library construction method for super-resolution ribosome profiling in Arabidopsis. Plant Methods. 2022;18:115.
80. Skowronek P, Thielert M, Voytik E, Tanzer MC, Hansen FM, Willems S, et al. Rapid and in-depth coverage of the (Phospho-)proteome with deep libraries and optimal window design for dia-PASEF. Mol Cell Proteomics. 2022;21(9):100279.
81. Demichev V, Messner CB, Vernardis SI, Lilley KS, Ralser M. Dia-nn: neural networks and interference correction enable deep proteome coverage in high throughput. Nat Methods. 2020;17:41–4.
82. Gerault MA, Camoin L, Granjeaud S. Diagui: a shiny application to process the output from DIA-NN. Bioinformatics Adv. 2024;4:vbae001.
83. Li Y. IsoSeq3 - scalable de novo isoform discovery from single-molecule PacBio reads. 2018.
84. Biosciences P. PacBio Secondary Analysis Tools on Bioconda. 2020. https://github.com/PacificBiosciences/pbbioconda.

85. Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 2019;20:275.
86. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015;6:11.
87. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
88. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
89. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. Bioinformatics. 2016;32:767–9.
90. Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. 2005;33:W465-467.
91. Consortium TU. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2022;51:D523–31.
92. Li H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics. 2016;32:2103–10.
93. Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. Whole-genome annotation with BRAKER. Methods Mol Biol. 2019;1962:65–95.
94. Gabriel L, Hoff KJ, Bruna T, Borodovsky M, Stanke M. TSEBRA: transcript selector for BRAKER. BMC Bioinformatics. 2021;22:566.
95. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 2003;31:5654–66.
96. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8:1494–512.
97. Shumate A, Salzberg SL. Liftoff: accurate mapping of gene annotations. Bioinformatics. 2021;37:1639–43.
98. Dainat J, Hereñú D, Murray K, Davis E, Crouch K, LucileSol, Agostinho N, Pascal-Git, Zollman Z. AGAT: Another GFF analysis toolkit to handle annotations in any GTF/GFF format. Zenodo Version v0.8.0. 2023.
99. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. Interproscan 5: genome-scale protein function classification. Bioinformatics. 2014;30:1236–40.
100. Galaxy C. The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. Nucleic Acids Res. 2024;52:W83–94.
101. Altenhoff AM, Levy J, Zarowiecki M, Tomiczek B, Warwick Vesztrocy A, Dalquen DA, et al. OMA standalone: orthology inference among public and custom genomes and transcriptomes. Genome Res. 2019;29:1152–63.
102. Train CM, Pignatelli M, Altenhoff A, Dessimoz C. lHam and pyHam: visualizing and processing hierarchical orthologous groups. Bioinformatics. 2019;35:2504–6.
103. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
104. Su J, Zheng Z, Ahmed SS, Lam TW, Luo R. Clair3-trio: high-performance Nanopore long-read variant calling in family trios with trio-to-trio deep neural networks. Brief Bioinform. 2022. https://doi.org/10.1093/bib/bbac301.
105. Mattick J. pbsv. 2022. https://github.com/PacificBiosciences/pbsv.
106. Nikolayeva O, Robinson MD. edgeR for differential RNA-seq and ChIP-seq analysis: an application to stem cell biology. Methods Mol Biol. 2014;1150:45–79.
107. Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat Methods. 2016;13:731–40.
108. Perez-Riverol Y, Bai J, Bandla C, Garcia-Seisdedos D, Hewapathirana S, Kamatchinathan S, Kundu DJ, Prakash A, Frericks-Zipper A, Eisenacher M, et al. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. Nucleic Acids Res. 2022;50:D543–D552.
109. Zhong Z, Bailey M, Kim Y, Afsharyna NP, Parker B, Arathoon L, Li X, Rundle CA, Behrens A, Nedialkova D, Slavov G, Hassani-Pak K, Lilliey KS, Theodoulou FL, Mott R. The distinct roles of genome, methylation, transcription, and translation on protein expression in Arabidopsis thaliana resolve the Central Dogma's information flow. ProteomeXchange Consortium, PRIDE partner repository dataset PXD058342. 2025.
110. Mott R. The distinct roles of genome, methylation, transcription, and translation on protein expression in Arabidopsis thaliana resolve the Central Dogma's information flow, software repository. Zenodo. 2025. https://www.zenodoorg/records/15467917.

Publisher's Note