PlantCAD Information Content Weighted Nucleotide Probability Tracks (B73 v5 ± 5 kb)
Category → AI and Machine Learning

This set of eight browser tracks visualizes how strongly each base is *expected* at every position surrounding all annotated B73 gene models, as inferred by **PlantCaduceus _l32 (PlantCAD)**. PlantCAD is a plant-specific DNA language model trained with Caduceus/Mamba architectures on 16 flowering-plant genomes and shown to transfer accurately to maize despite 160 million years of divergence (see paper: https://doi.org/10.1101/2024.06.04.596709 and GitHub: https://github.com/kuleshov-group/PlantCaduceus).

For every nucleotide in a window from -5 kb upstream to +5 kb downstream of each gene model:

1. Mask the focal base,
2. Ask PlantCAD to predict the masked token (“zero-shot”),
3. Retrieve the probability assigned to **A, C, G, T**,
4. Convert the four-way entropy to **information content (IC)** on a 0–2 bit scale,
5. Multiply the per-base probability by IC (i.e., *P(base)* × *IC*) to emphasize positions that are both confidently predicted and information-rich.
6. Convert the data to wig/Bigwig

Each specific track description:

- **PlantCAD IC-Weighted, nucleotide A:** This track shows a score for each position in the genome that highlights regions rich in the nucleotide Adenine (A) and where PlantCAD has high confidence in its prediction. Higher scores may indicate the presence of important regulatory elements, such as TATA boxes.
- **PlantCAD IC-Weighted, nucleotide C:** This track shows a score for each position in the genome that highlights regions rich in the nucleotide Cytosine (C) and where PlantCAD has high confidence in its prediction. Higher scores may indicate the presence of important features, such as CpG islands or GC-boxes.
- **PlantCAD IC-Weighted, nucleotide G:** This track shows a score for each position in the genome that highlights regions rich in the nucleotide Guanine (G) and where PlantCAD has high confidence in its prediction. Higher scores may indicate the presence of important features, such as G-quadruplexes or GC-boxes.
- **PlantCAD IC-Weighted, nucleotide T:** This track shows a score for each position in the genome that highlights regions rich in the nucleotide Thymine (T) and where PlantCAD has high confidence in its prediction. Higher scores may indicate the presence of important regulatory elements, such as TATA boxes.
- **PlantCAD IC-Weighted, nucleotide A/C/G/T summary*:** This track provides a combined view of the individual A, C, G, and T tracks. At each position, it highlights the nucleotide with the highest score, indicating the most likely and confidently predicted nucleotide. This allows for easy comparison and identification of the strongest signals across all four nucleotides.
- **PlantCAD Information Content (0–2 bits):** This track displays a measure of PlantCAD's certainty at each position. High values (close to 2) indicate PlantCAD is

confident in its nucleotide prediction, suggesting these regions might be functionally important and less likely to vary. Low values (close to 0) indicate PlantCAD is uncertain, suggesting these regions might be more flexible or variable. This can be helpful for understanding the potential impact of genetic variations.

- **PlantCAD IC-Weighted Reference Allele:** This track shows a score for each position that highlights regions where PlantCAD predicts the nucleotide should match the standard maize reference genome (B73) with high confidence. Higher scores indicate positions where PlantCAD strongly expects the reference allele to be present.
- **PlantCAD IC-Weighted Top Alternate Allele:** This track highlights positions in the genome where PlantCAD strongly predicts a nucleotide that is *different* from the standard maize reference genome (B73). The score at each position reflects both the likelihood of this alternate nucleotide and PlantCAD's confidence in this prediction. Higher scores on this track help identify potential genetic variations (SNPs or indels) that might have a functional impact.
- **PlantCAD Allelic Information Content Shift (-2 to 2 bits):** A track showing the information-content-weighted probability difference between the most likely alternate allele and the B73 reference allele at each base (ΔIC×P), so that positive values highlight where the alternate is favored and negative values where the reference remains are stronger.

Recommendations: Overlaying these PlantCAD tracks with other types of data, such as information about DNA methylation, chromatin structure, known genetic variants, or regulatory motifs, can help identify important regulatory regions, support gene model annotations, pinpoint acceptor/donor loci, prioritize genetic variants, and discover conserved DNA sequences.

***Note on A/C/G/T summary Zoom Levels:** As you zoom out, the data displayed transitions from individual raw scores at each locus to values averaged across multiple loci. Consequently, the scores at broader views may appear lower than individual peaks, but this averaging effectively illustrates the general trends in nucleotide prevalence across wider genomic regions.

MaizeGDB 2024: Signed log-odds Minor-Allele Frequency (MAF)

Summary

This quantitative track shows the **signed log<sub>10</sub> odds-ratio of minor-allele frequency** (score = log<sub>10</sub>[MAF / (0.05)]) at every polymorphic site detected in MaizeGDB 2024 – High Quality dataset which covers ~75 millions variant sites from ~1,500 lines (inbreds, landraces, teosintes). Minor Allele Frequency is the proportion of the less common allele at a genetic locus within a given population.

- **Positive bars (blue)** → common variants (MAF > 0.05)
- **Negative bars (red)** → less common and rare variants (MAF < 0.05)
- **Zero baseline** → variants at or near 5% in this population or no variant found at this locus

Because the score is symmetric around zero, the track visually separates loci that *lack* variation from those that have **very rare alleles** (deep red) or **nearly fixed alternative alleles** (tall blue).

Table to interpret the scores

MAF	**Score = log10(MAF / 0.05)**	**Bins (Percent and number of accessions with the alternative allele)**
0.5	1	Very common (50%, ~ 750 accessions)
0.25	0.7	Common (25%, ~375 accessions)
0.1	0.3	Low-frequency (10%, ~150 accessions)
0.05	**0**	Pivot/frequency boundary (5%, ~75 accessions)
0.01	-0.7	Rare (1 %, ~15 accessions)
0.005	-1	Very rare (0.5 %, < 10 accessions)
0.001	-1.6	Ultra-rare (0.1 %, 2 accessions)
0.0006	-1.9	Homozygous singletons (.06%, 1 accession)
0.0003	-2.2	Heterzygous singleton (.006%, 1 accession)

File provenance & citations

- **Reference assembly:** *Zea mays* B73 RefGen_v5
- **Variant calling & annotation:** as described in MaizeGDB 2024 (https://doi.org/10.1093/g3journal/jkae281).

Please cite the original data releases when using these tracks in publications.