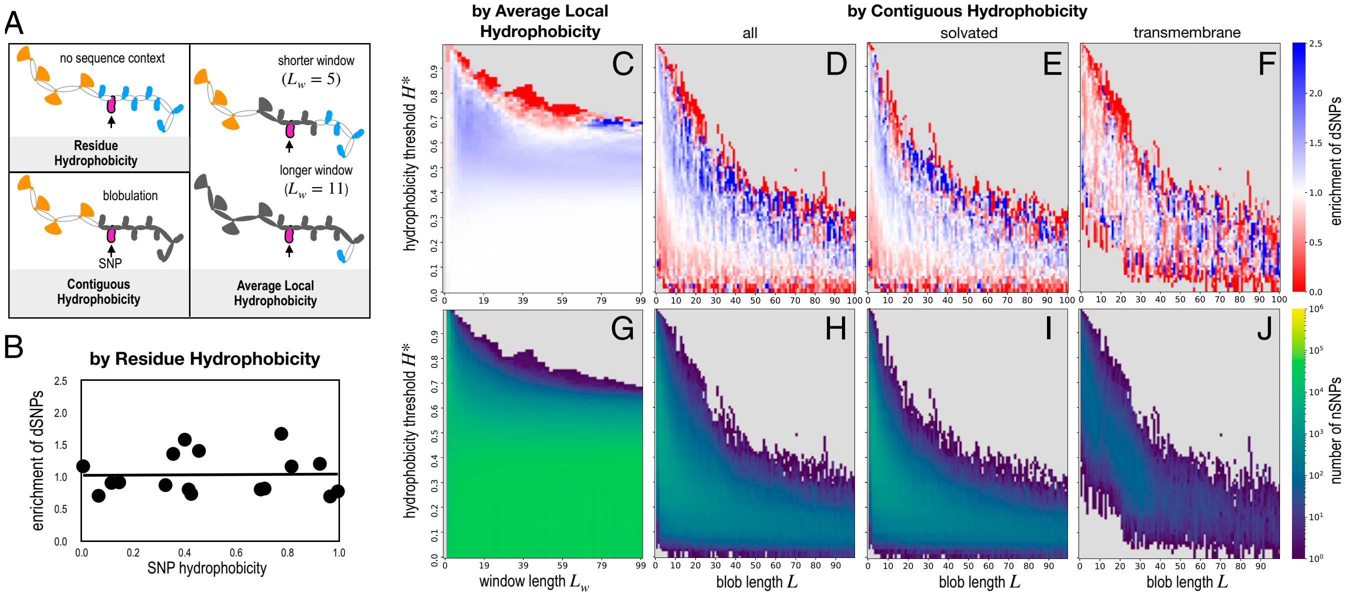

Fig. 2. Effect of segmentation approach, length, hydrophobicity threshold, and solvation on calculated enrichment of dSNPs in hydrophobic segments. (*A*) Illustration of three measures of SNP hydrophobicity (residue, contiguous, and average local) for the indicated SNP, found within a hypothetical peptide chain composed of residues classified as hydrophobic (blue ovals) or nonhydrophobic (orange fans) for a given $H^\star$, as in Fig. 1*B*. Unconstrained-length blobulation determines the local sequence (shaded in gray) by detecting contiguous hydrophobic residues, which together form an h-blob of length L; the moving-window approach determines the local sequence using a fixed number L_w of residues centered around the SNP. (*B*) Enrichment of dSNPs relative to nSNPs as a function of hydrophobicity of the reference allele, with line of best fit. No trend or significant correlation is observed (Pearson's $r = 0.02$, $P = 0.94$, $n = 17$). (*C*–*F*) Enrichment of dSNPs in hydrophobic segments, as a function of segment length and threshold, for (*C*) fixed-length hydrophobic windows of length L_w in which the average hydrophobicity is above $H^\star$, and (*D*–*F*) h-blobs of length L, calculated with the threshold $H^\star$ for (*D*) all SNPs, (*E*) those outside of transmembrane domains, and (*F*) those in transmembrane domains. (*G*–*J*) The total number of nSNPs per bin for the corresponding enrichment heatmaps in *C*–*F* (e.g., *I* shows the nSNP counts for the enrichments in *E*). Each panel (*C*–*J*) is colored according to the scale at *Right* end of the row, and bins with no data are colored gray.

$n = 17$) between lone SNP hydrophobicity and dSNP enrichment, meaning that the hydrophobicity of a residue considered in isolation does not show this particular signature of functionality.

The effect of average local hydrophobicity on the enrichment of dSNPs was calculated using moving windows of length L_w centered around each SNP. While there is no "standard" window size, most SNP prediction programs use a window size in the range of 1 to 21 residues (16, 20, 37, 38). The window size is chosen to balance concerns that small window sizes may not accurately capture the "local" sequence (39–41) whereas large window sizes can decrease the signal-to-noise ratio (42). Here we computed the mean window hydrophobicity $\bar{H}_i$ for all SNPs i in our SNP dataset, while also varying L_w. Fig. 2*C* shows the enrichment of dSNPs for which $\bar{H}_i > H^\star$, for the range of moving-window widths $L_w = 1$ to 99. As is evident in Fig. 2*C*, the enrichment of dSNPs is relatively insensitive to the window size for the regime where $L_w \geq 6$ and $H^\star \leq 0.65$. The total count of nSNPs in each bin is shown in Fig. 2*G* and was similarly insensitive to window size for larger thresholds. These results suggest that distant residues introduce noise that averages out in a proteome-wide analysis, but their inclusion in the window would still reduce precision for any individual SNP.

Surprisingly, we detect a narrow band of dSNP depletion for windows with a high average hydrophobicity (the red signal for $H^\star \geq 0.65$ in Fig. 2*C*). This signal is due to only a handful of SNPs (see the counts in Fig. 2*G*), but we observe a similar pattern using the blobulation algorithm, and we discuss its origins in the next sections.

Contiguous Hydrophobicity Is a Strong Indicator of Disease Association. The use of a fixed-width window neglects the inherent dispersion in the size of protein modules, which are captured using blobulation (Fig. 2*A*). To quantify the blob properties for each SNP, we used a blobulation variant we call unconstrained-length blobulation, which fixes the threshold $H^\star$ and a reference residue i but imposes no minimum blob length. This approach first tests whether the hydropathy score for residue i exceeds $H^\star$, and if so, it calculates the exact length L of the h-blob that contains residue i. Unconstrained-length blobulation is formally equivalent to whole-sequence blobulation with $L_{\text{min}} = 1$, but is more efficient since we are analyzing only the relevant part of the sequence.

Specifically, we applied unconstrained-length blobulation to each SNP in the dataset (using the reference allele) and a given hydrophobicity threshold $H^\star$ and then determined L. We repeated this calculation for a series of $H^\star$ values, and for dSNPs and nSNPs separately (Dataset S3), and then tabulated the proportion of dSNPs and nSNPs in each $(H^\star, L)$ bin (Dataset S4). The resulting enrichment of dSNPs as a function of blob length L and threshold $H^\star$ is shown in Fig. 2*D*, and the total number of nSNPs per bin is shown in Fig. 2*H*. We observe a consistent relationship between hydrophobicity of the local blob and dSNP enrichment. dSNPs are depleted in weakly hydrophobic blobs, are neutral for moderately hydrophobic blobs, and become more enriched as the blob gets longer and/or satisfies a stricter hydrophobicity threshold. The trend is primarily monotonic, which supports the hypothesis that hydrophobic blobs constitute functional elements.

We do find an exception to the trend at the plot boundary: Blobs that satisfied the very strictest criteria were depleted in dSNPs, consistent with the results using the moving-window analysis. The depletion signal persisted even when bins with very few samples were removed. In the next section, we consider two potential reasons for this depletion: 1) dSNPs in these blobs are so deleterious that they are selected out of the population or 2) some subset of nSNPs is functional and under balancing selection or relaxed constraint. In addition to a consistent trend, the analysis returns a large spread in enrichment/depletion values: 3.2% of the bins in Fig. 2*D* are significantly depleted below 0.5, while