Laminar recordings in frontal cortex suggest distinct layers for maintenance and control of working memory

André M. Bastos[a,b,1], Roman Loonis[a,b,1], Simon Kornblith[a,b], Mikael Lundqvist[a,b], and Earl K. Miller[a,b,2]

[a]The Picower Institute for Learning and Memory, Massachusetts Institute of Technology, Cambridge, MA 02139; and [b]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139

Edited by Peter L. Strick, University of Pittsburgh, Pittsburgh, PA, and approved December 20, 2017 (received for review June 14, 2017)

All of the cerebral cortex has some degree of laminar organization. These different layers are composed of neurons with distinct connectivity patterns, embryonic origins, and molecular profiles. There are little data on the laminar specificity of cognitive functions in the frontal cortex, however. We recorded neuronal spiking/local field potentials (LFPs) using laminar probes in the frontal cortex (PMd, 8A, 8B, SMA/ACC, DLPFC, and VLPFC) of monkeys performing working memory (WM) tasks. LFP power in the gamma band (50–250 Hz) was strongest in superficial layers, and LFP power in the alpha/beta band (4–22 Hz) was strongest in deep layers. Memory delay activity, including spiking and stimulus-specific gamma bursting, was predominately in superficial layers. LFPs from superficial and deep layers were synchronized in the alpha/beta bands. This was primarily unidirectional, with alpha/beta bands in deep layers driving superficial layer activity. The phase of deep layer alpha/beta modulated superficial gamma bursting associated with WM encoding. Thus, alpha/beta rhythms in deep layers may regulate the superficial layer gamma bands and hence maintenance of the contents of WM.

cortical layers | oscillations | working memory | frontal cortex

Working memory (WM) is associated with neural activity during a memory delay. This is thought to be due to recurrent connections between columns of pyramidal neurons in superficial cortical layers (1), but support for this has been mixed. One study found that delay activity was shared across superficial and deep layer neurons (2), whereas another reported delay activity neurons ("late storage units") at more superficial depths (3). This uncertainty may be due to the previous use of single-contact electrodes, which make it difficult to assess the depth of the recorded signals.

Another question is whether the frontal cortex shows similar layer-specific properties as the visual cortex (4–7). In the visual cortex, multiple-contact "laminar" electrodes have revealed that gamma (>30 Hz) oscillations are prominent in superficial/middle layers, while slower oscillations (alpha/beta; 10–30 Hz) are prominent in deep layers (5, 7, 8). Deep layer alpha activity drives superficial alpha activity (4, 6), and the phase of deep layer alpha modulates the amplitude of superficial layer gamma (9). Similar tests in the supplemental eye fields have reported gamma power in superficial layers (10, 11); however, one of those studies failed to find evidence of deep layer low-frequency oscillations coupled with superficial gamma (11), leading to the conclusion that frontal cortex laminar dynamics might be fundamentally different. Neither study examined WM-related activity in the frontal cortex with laminar electrodes.

We recorded both spiking and local field potential (LFP) activity with multilaminar electrodes in six frontal cortex areas (PMd, 8A, 8B, SMA/ACC, DLPFC, and VLPFC) in three WM tasks. This revealed delay period spiking predominately in superficial layers and laminar dynamics in frontal cortex similar to those in visual cortex, suggesting a layer-specific pattern of recurring dynamics between visual and frontal cortices.

Results

Gamma Power Peaks in Superficial Layers and Alpha/Beta Peaks in Deep Layers. Three monkeys performed three different WM tasks (Fig. 1 *A–C*). Either spatial or object identity information had to be retained during a delay period. Laminar probes with spacing of 100–200 μm between contacts recorded LFPs and neuronal spiking from all cortical layers (Fig. 1 *D–F*). They were lowered as perpendicular as possible to the cortex to ensure an even sampling of the different layers. We completed a total of 60 U/V probe recordings in frontal cortex (Fig. 1*G* and *SI Appendix*, Table S1). The middle cortical layer (bottom of layer 3/layer 4) was identified, using current source density (CSD) analysis, by the presence of a current sink in response to a visual stimulus (see *SI Appendix, Experimental Procedures*). We aligned all of the data from all electrodes to the middle layer (i.e., the contact with the first significant CSD sink; see *SI Appendix*, Fig. S1 for the average CSD profiles).

We calculated power from 200 ms before the visual stimulus to 500 ms post stimulus. High frequency power peaked in superficial layers and low frequencies peaked in deep layers. Fig. 1*H* plots examples of low- and high-passed LFPs from one laminar recording, and Fig. 1*I* is a power spectrum, illustrating that high frequencies (above ∼40 Hz) dominated in superficial layers and low frequencies (below ∼20 Hz) dominated in deep layers. To quantify these differences in power across each laminar probe, we normalized power at each frequency (1–500 Hz) and each contact by the maximal power at that frequency across contacts. For each frequency and each session, the contact with maximal

Significance

The anatomy and dynamics of different layers of the cerebral cortex are distinct. Physiological work in the sensory cortex has investigated how different layers process sensory inputs, and how they are engaged during attention tasks. In the frontal and prefrontal cortices, where lamination is present, very few studies have investigated the role of distinct layers for cognition. We studied frontal cortex laminar neuronal activity as monkeys performed working memory tasks. Spiking and gamma-band activity (50–150 Hz) in the superficial layers reflected active maintenance of working memories. Alpha/beta frequencies (4–22 Hz) in the deep layers modulated the gamma activity in the superficial layers. This might serve a control function, allowing information to enter or exit active storage in superficial layers.

Author contributions: A.M.B., R.L., and E.K.M. designed research; A.M.B., R.L., and S.K. performed research; A.M.B., R.L., and M.L. analyzed data; A.M.B., R.L., M.L., and E.K.M. wrote the paper; and E.K.M. supervised the study.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]A.M.B. and R.L. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: ekmiller@mit.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1710323115/-/DCSupplemental.

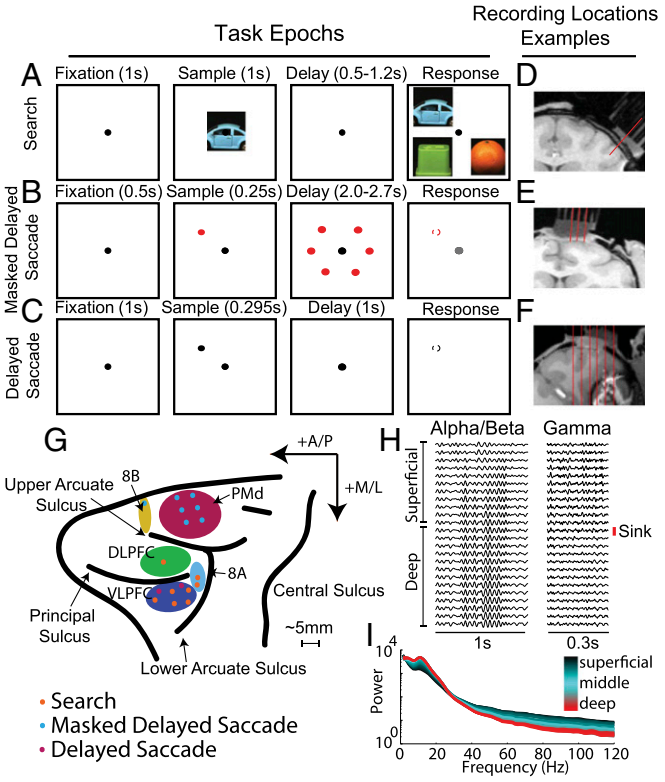


Fig. 1. (*A*) Visual search task. A match between sample and test image was chosen after a delay (0.5–1.2 s) by making a saccade to the match. Each image was positioned randomly at any one of four possible locations (*Upper Right*, *Lower Right*, *Upper Left*, and *Lower Left*). (*B*) Masked delayed saccade. After a sample period, during which a single spatial location was cued (one of six possible locations), the animal had to hold fixation through a variable delay (2.2–2.7 s) and the presentation of the visual mask. After this delay, and when the fixation point color changed, the animal had to saccade to the previously cued location. (*C*) Delayed saccade. After a sample period, during which a single spatial location was cued (one of four possible locations), the animal had to hold fixation through a fixed delay (0.99 s) and saccade to the cued location when the fixation dot disappeared. (*D* and *E*) The small red lines indicate sample trajectories chosen to be as perpendicular as possible to cortex. (*F*) The small red lines indicate sample trajectories that were possible given the recording hardware. Only the third trajectory from the left was used for laminar recordings. (*G*) We recorded across frontal cortex. The different colored dots indicate the task, and the letters indicate the corresponding anatomic region. In addition to those labeled, we recorded from the anterior cingulate cortex (ACC) and the supplementary motor area (SMA) (*SI Appendix*, Fig. S12). (*H*) Sample LFP recordings were bandpass-filtered at 10–25 Hz (*Left*) and 40–160 Hz (*Right*). The red line marks the location of the first significant CSD sink and the border between the superficial and deep layers. (*I*) A sample power spectrum with a clear alpha/beta bump (between 10 and 25 Hz) and broadband gamma (>40 Hz). The variations across layers are plotted as a color gradient (black, superficial to red, deep).

power had a value of 1 and other contacts had values relative to this maximum.

Fig. 2*A* shows the mean power profile across the entire dataset. The middle cortical layer is at depth 0 (where the early sink was detected), negative depths are superficial layers (layers 1–3), and positive depths are deep layers (layers 5–6). Red colors indicate the maximal power and blue the minimal. The superimposed black line is the mean depth at which the maximal power occurred for each frequency. We tested whether each frequency consistently peaked away from zero. Lower frequencies (4–22 Hz) had their maximal power at sites significantly deeper than depth zero, while a continuous band of higher frequencies (58–260 Hz) had their maximal power above zero, in superficial layers (sign test across sessions, Bonferroni corrected for multiple comparisons, $P < 0.05$).

We collapsed Fig. 2*A* into two separate profiles by averaging across the alpha/beta (4–22 Hz) and gamma (58–260 Hz) frequency bands (Fig. 2*B*). The peak gamma power (red line) occurred in superficial layers, 400 μm above the sink, and the peak alpha/beta power (blue line) occurred in deep layers, 600 μm below the sink. The cross-over point between the profiles (the intersection of the blue and red lines) occurred between −100 and −200 um, nearly identical to the location of the CSD sink. Thus, gamma power was prominent in superficial layers, and alpha/beta in deep. In middle layers (from 200 μm above to 300 μm below the sink), there was a transition zone in which neither gamma nor alpha/beta predominated. These results were present in each of the three tasks (Fig. 2 *C*–*E*) and in all of the individual areas that we sampled (*SI Appendix*, Fig. S2), with the exception of the alpha/beta profile for one area (8B), which was qualitatively similar. This pattern was also observed when we defined the middle layer based on CSD analysis time locked either to a monitor screen flash or to sample onset during the WM task (*SI Appendix*, Fig. S1).

Delay Period Activity in Superficial Layers. WM has been linked to persistent modulation of delay period spiking activity, but whether this activity is layer-specific is unclear. We found that delay period spiking was largely localized to superficial layers. To measure multiunit activity (MUA), we used rectified, high-pass signals >500 Hz. To assess delay period modulation, we took the absolute value of the mean change in MUA between the delay and the baseline (the presample fixation window), and *z*-scored it by the SD of delay period MUA across trials (*SI Appendix, Experimental Procedures*). This normalization step ensured that differences in the overall strength of MUA across sessions were deemphasized before pooling. We used the absolute values because activity in delays could increase or decrease relative to baseline (12).

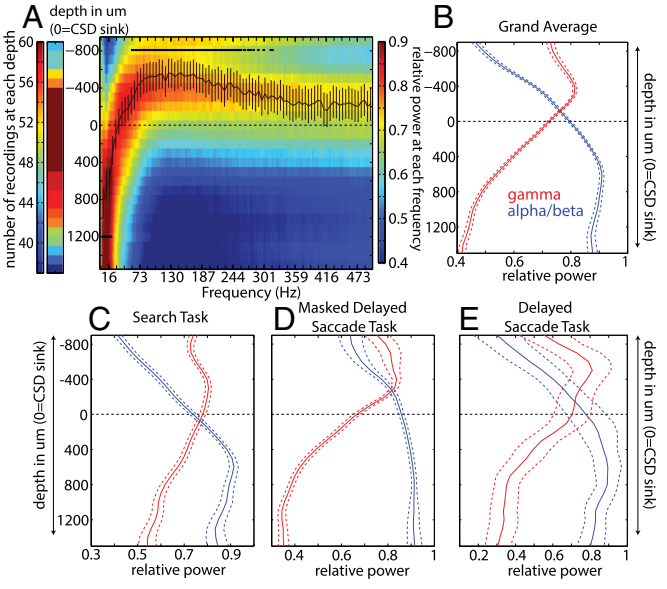


Fig. 2. (*A*) Number of LFP recordings performed per depth (*Left*) and normalized power averaged across multicontact probes with respect to depth and frequency (*Right*). Red indicates greater power at a particular depth; blue, less power. The black line represents the average depth at which the power at each frequency peaks. Error bars ± 1 SEM. The black stars indicate frequency bins at which the mean depth was significantly different from zero (Bonferroni- corrected for multiple comparisons). (*B*) Normalized power averaged across low (4–22 Hz, blue line) and high (50–250 Hz, red line) frequencies. Error bars ± 1 SEM. (*C*–*E*) Normalized power profiles across low and high frequencies for each task. Error bars ± 1 SEM.

Fig. 3 *A*, *C*, and *E* shows the time-resolved delay period MUA modulation averaged across sessions for each task. In the figure, red indicates a change in MUA from the presample fixation baseline, and blue indicates little or no change. In all tasks, the greatest change in MUA was in superficial layers. Fig. 3 *B*, *D*, and *F* illustrates the time-averaged delay period modulation. It shows that the greatest modulations were in superficial layers. Averaging across all tasks, delay MUA peaked at 400 μm above the sink and dropped in deeper cortical layers ($P < 0.002$, sign test across sessions; Fig. 3*G*). This was the same depth at which gamma power peaked in the LFP (Fig. 2*B*). Moreover, the laminar profile of the average modulation of delay period MUA was positively correlated (Spearman rank correlation across depth from −900 to 1,500 μm, $\rho = 0.84$; $P = 2E\text{-}6$) with the laminar profile of gamma power (compare with Fig. 2*B*) and negatively correlated with alpha/beta power (Spearman rank correlation, $\rho = -0.67$; $P = 3E\text{-}4$). This correlation was present in each task individually (Spearman rank correlation, visual search; gamma: $\rho = 0.79$, $P < 1E\text{-}5$; alpha/beta: $\rho = -0.57$, $P = 0.003$; masked delayed saccade, gamma: $\rho = 0.80$, $P < 1E\text{-}5$, alpha/beta: $\rho = -0.66$, $P < 0.001$; delayed saccade, gamma: $\rho = 0.86$, $P < 1E\text{-}5$, alpha/beta: $\rho = -0.68$, $P < 0.001$). In addition, the peak delay period MUA was in superficial layers for all six areas ($P = 0.03$, sign test; *SI Appendix*, Fig. S3).

The rectified, high-pass signal that we used to measure MUA ("analog MUA") is thought to capture the mean of all spiking activity within the local vicinity of the recording contact (13). However, we noted some spectral overlap between this signal and the gamma power in superficial layers, raising the possibility that the greater gamma power was due to the MUA (or vice versa). To address this, we used a thresholded signal to measure the spike rate. This signal was more conservative than the MUA signal, capturing only small groups of units with large spikes near the contact (hereinafter referred to as "units"). Representative spike waveforms and firing rates for a single session are shown in *SI Appendix*, Fig. S4. We identified a total of 423 units.

Analysis of delay period modulation based on spike rates confirmed that the proportion of units with delay activity was higher in superficial layers compare with deep layers (53% vs. 34%; $P = 3E\text{-}4$, χ^2 proportion test; *SI Appendix*, Fig. S5*B*, *Inset*). Spiking during the delay period carried significant information about the sample, as measured by percent explained variance (PEV) (*SI Appendix, Experimental Procedures* and Fig. S6). In addition, the peak spike PEV value was observed in superficial layers at a depth of −400 um, the same depth at which gamma power peaked. The greater proportion of modulated units in superficial layers was not the result of a poor signal-to-noise ratio or lack of units in deep layers. In fact, baseline firing rates were higher in the deep layers (*SI Appendix*, Fig. S5*C*), as were unit

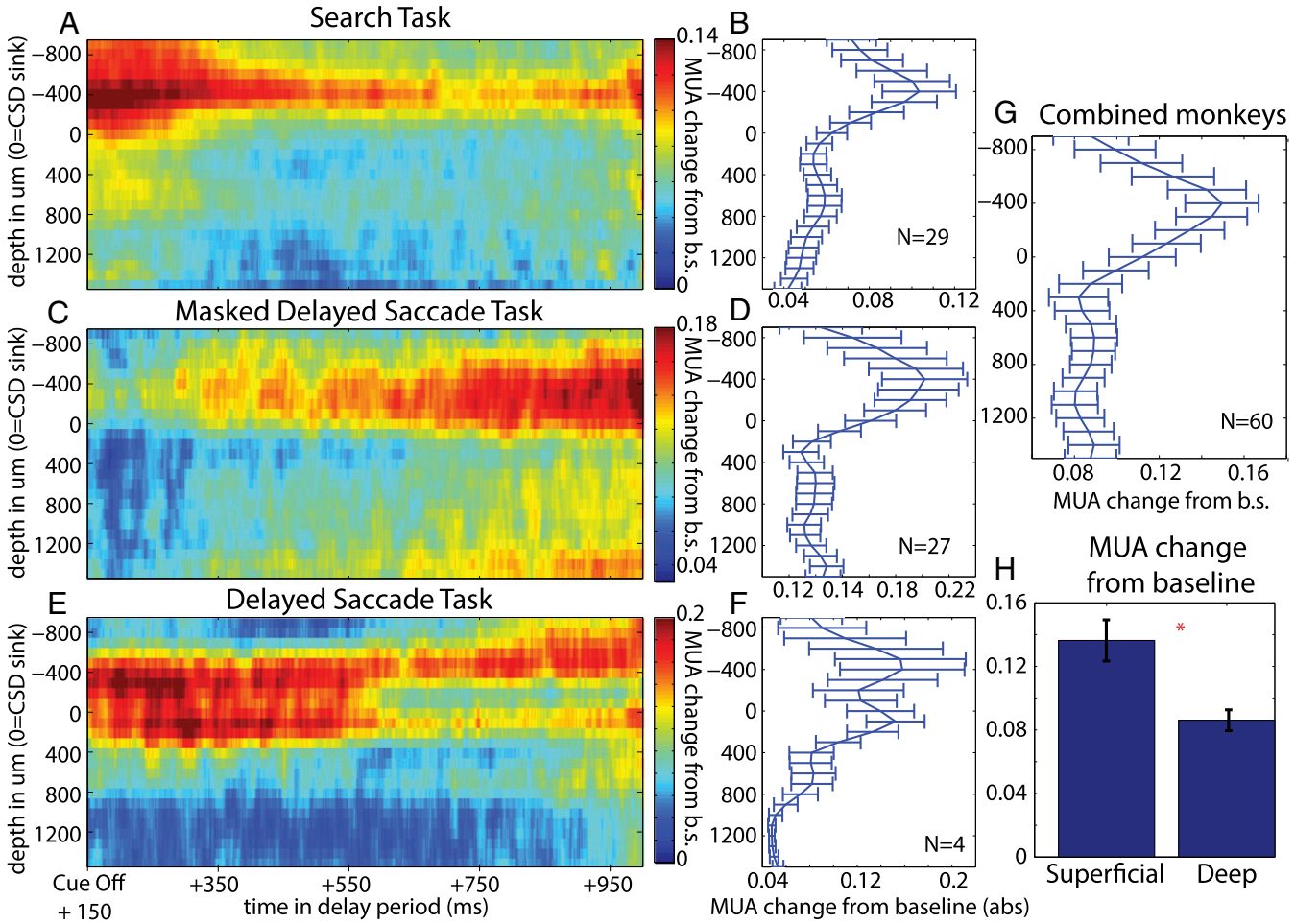


Fig. 3. (*A*, *C*, and *E*) Delay period, MUA modulation across cortical depth and time, plotted from 150 ms after sample offset to 1 s into the delay, for the visual search, masked delayed saccade, and delayed saccade tasks. (*B*, *D*, and *F*) The mean delay period MUA modulation per cortical depth for each task. Error bars ± 1 SEM. (*G*) The delay period MUA modulation averaged across all tasks per cortical depth. Error bars ± 1 SEM. (*H*) The mean MUA modulation averaged across all tasks, and all superficial or deep contacts. Error bars ± 1 SEM.

yields (*SI Appendix*, Fig. S5*D*). A similar delay period profile was seen when we aligned depth based on the transition between cerebrospinal fluid and the gray matter (*SI Appendix*, Fig. S7). Finally, MUA in superficial layers did not dominate all task periods. During sample processing, MUA activity was more prominent in middle layers (in granular areas), and during saccade generation, both deep and superficial layers became active (*SI Appendix*, Fig. S8).

Gamma Bursts in Superficial Layers Encode Stimulus Information During the Delay. Recent work has shown that oscillatory gamma bursting in the prefrontal cortex (PFC) is associated with encoding of stimuli in WM (14). We tested for its layer-specificity. To distinguish this from the power analyses described above, we defined bursts as periods in which the power in the alpha/beta (4–22 Hz) and gamma (50–150 Hz) bands exceeded the mean power at each frequency band by 2 SD for three oscillation cycles (*SI Appendix, Experimental Procedures*). The gamma and alpha/beta burst rates during the delay and presample fixation baseline periods are shown in Fig. 4 *A* and *C*. The average gamma burst rate increased during the delay relative to the baseline ($P = 0.004$, sign test; Fig. 4*A*), and the alpha/beta burst rate decreased ($P < 1E\text{-}8$, sign test; Fig. 4*C*).

We tested whether delay period gamma and alpha/beta bursts carried information about which cue was held in WM by calculating the PEV between the burst rate and cued object/location during the delay (*SI Appendix, Experimental Procedures*). Fig. 4*B* shows the profile of information in gamma bursts by layer, and Fig. 4*D* for alpha/beta bursts. Over sessions, gamma bursting was more informative in superficial layers than in deep layers ($P = 0.004$, sign test over sessions; Fig. 4*B*, *Inset*). Furthermore, the amount of gamma bursting information per layer was strongly correlated with gamma power (Spearman's rank correlation, $\rho = 0.82$; $P < 3E\text{-}6$). Information from alpha/beta bursting was weaker, as might be expected given the very low burst rate during the delay. Information in alpha/beta bursting was not significantly different between deep and superficial layers, but trended toward an increase in deep layers (Fig. 4*D*). The amount of information in alpha/beta bursting correlated with the laminar alpha/beta power profile ($\rho = 0.51$; $P < 0.02$).

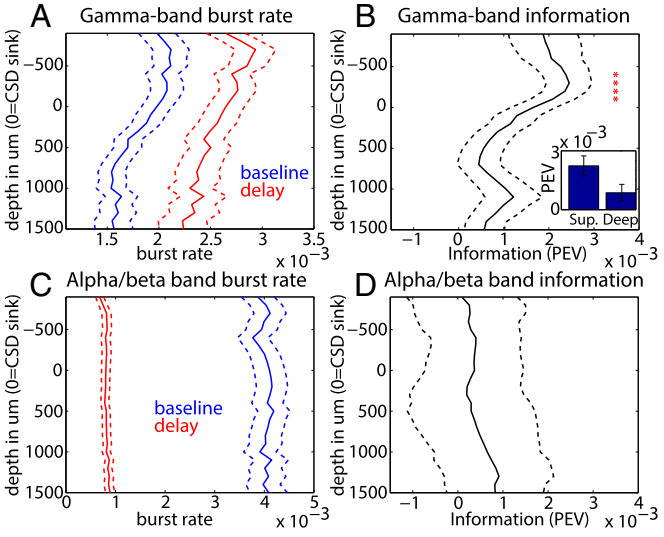


Fig. 4. (*A*) Gamma burst rates at baseline (blue) and during the delay (red). (*B*) Percentage of explained variance (omega-squared) of the gamma bursts across different cortical depths. Red asterisks indicate depths at which there is significantly nonzero PEV across sessions ($P < 0.05$, Bonferroni-corrected). (*Inset*) Mean PEV across all superficial and deep contacts, respectively. Error bars ± 1 SEM. (*C*) Same as *A*, but for alpha/beta. (*D*) PEV of alpha/beta bursts across different cortical depths. Error bars ± 1 SEM.

Alpha/Beta Oscillations in Deep Layers Modulate Superficial Layers, and Coupling Is Reduced During the Delay. We next tested for interactions between oscillatory activity and layers. To assess frequency-resolved directed interactions, we applied nonparametric Granger causality (GC; a measure of statistical prediction between time series; *SI Appendix, Experimental Procedures*) analysis to LFPs within superficial and deep layers. We found that the GC spectrum had peaks in the alpha/beta range, and that directed interactions were asymmetric. Deep layer alpha/beta drove superficial layer alpha/beta more than the other way around (4–22 Hz, $P = 0.002$, sign test; Fig. 5*A*).

GC is a linear measurement of interlaminar interactions, in the sense that interactions are only tested between different channels at the same frequency. To test for cross-frequency (nonlinear) interactions, we next investigated cross-frequency coupling (CFC) between superficial and deep layers. To test whether the phase of the slower frequency band (alpha/beta band) coupled with the amplitude of the higher-frequency band (gamma band), we used the modulation index (15), a measure of how nonuniformly distributed the amplitudes of one frequency band are across the phase space of another (*SI Appendix, Experimental Procedures*). We systematically calculated CFC at every possible combination of (alpha/beta) phase-providing channel and (gamma) amplitude-providing channel.

We found that the influence of the deep alpha/beta phase coupled with superficial gamma amplitude was stronger in the ascending (deep to superficial) direction than in the reverse direction ($P < 1E\text{-}4$, sign test over sessions; Fig. 5 *B* and *C*). This laminar profile of deep layer alpha/beta phasic modulation of gamma power in superficial layers was largely preserved across all six cortical areas (*SI Appendix*, Fig. S9). The GC and CFC results were basically consistent, with both indicating an ascending (deep to superficial) direction of influence or modulation.

To test whether deep to superficial coupling was task-modulated, we tested whether CFC and GC changed during presample fixation baseline vs. delay. CFC was significantly reduced during delay relative to baseline ($P < 0.0005$, sign test over sessions; Fig. 5*D* and *SI Appendix*, Fig. S10). In contrast, we found no significant differences in GC influence in the delay vs. baseline contrast ($P > 0.2$).

To test whether the interlaminar coupling between gamma and alpha/beta was excitatory or inhibitory, we performed power-power correlation analysis between all possible combinations of layers (*SI Appendix, Experimental Procedures*). Deep layer alpha/beta power was negatively correlated with superficial gamma power, consistent with an inhibitory influence (Fig. 5*E*). Power-power correlation in this compartment was significantly more negative than deep gamma power correlations with both deep and superficial alpha/beta power ($P < 0.05$, sign test over sessions; Fig. 5*F*). This suggests that deep layer alpha/beta regulates superficial alpha/beta (via ascending GC influence), which in turn regulates superficial layer gamma (via both CFC and negative power correlation).

Discussion

Evidence for Recurring Dynamics. Superficial and deep layers of frontal cortex exhibited distinct dynamics. Gamma power peaked in superficial layers, while alpha/beta power peaked in deep layers. The phase of these deep-layer alpha/beta oscillations modulated the amplitude of superficial gamma. Delay-period activity peaked in superficial layers. These dynamics were consistent across six distinct cortical areas (*SI Appendix*, Figs. S2, S3, and S9) spanning from premotor cortex to prefrontal cortex in our dataset, and match closely with reported results from visual cortex (4–9). The consistency and the specificity of these physiological effects suggest an underlying pattern of recurring neuronal dynamics shared between visual and frontal cortices (16, 17). Along with these consistencies, we also observed some qualitative deviations from these patterns, possibly as a result of comparing different tasks. Moreover, although many aspects of the dynamics were shared, this does not imply that each laminar

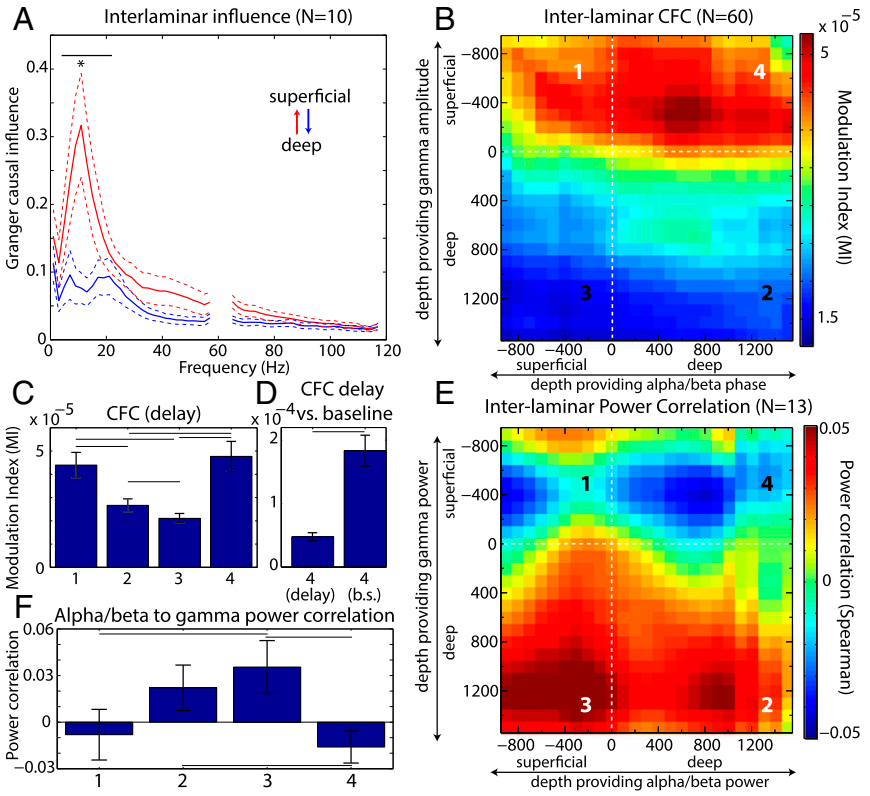


Fig. 5. (*A*) The Granger causal (GC) influence across frequency during the delay period. The red line is the GC of deep to superficial layers, and the blue is the reverse. Only sessions in which two laminar probes were placed within 2–4 mm of one another were used (see *SI Appendix, Experimental Procedures*). (*B*) Cross Frequency Coupling (CFC) between the phase of alpha/beta oscillations and the amplitude of gamma oscillations. Plotted across both axes is the CFC between specific cortical depths. (*C*) The mean CFC across all four possible laminar combinations during the delay-period: superficial phase to superficial amplitude (*Left*, 1), deep phase to deep amplitude (*Middle-Left*, 2), superficial phase to deep amplitude (*Middle-Right*, 3), and deep phase to superficial amplitude (*Right*, 4). (*D*) CFC between deep phase to superficial amplitude during the delay (*Left*) and the baseline (*Right*). (*E*) Correlation map between the power of alpha/beta and gamma over the delay period. Only recordings with a fixed delay period were used (four recordings from the delayed saccade task, nine from the search task). (*F*) Average power correlation between alpha/beta and gamma for all four possible laminar combinations. Same numbering scheme for laminar combinations as in *C*. Error bars ± 1 SEM. Lines indicate significantly different comparisons at $P < 0.05$.

circuit is identical. Therefore, further studies are needed to determine to what extent and in what areas these laminar dynamics are shared, and the functional consequence of this.

Delay Activity in Superficial Layers. Delay period spiking and gamma bursting encoding the contents of WM were most prominent in superficial layers. The co-occurrence of these two phenomena is consistent with reports that gamma bursts are associated with spiking that encodes a stimulus in WM (14). The relationship between layer 2/3 gamma and WM was also predicted by a WM model (18), which itself was based on known superficial layer connectivity and sparse activity patterns (1, 19, 20). The broadband nature of the average power spectrum in the gamma range (which lacked a clear peak) does not necessarily imply the lack of an oscillatory phenomenon, which could manifest as bursts of varying frequency within individual trials (14). Indeed, we found gamma bursts that increased in the delay and carried stimulus-related information, especially in superficial layers. The lower baseline firing rates in superficial layers, together with higher delay period information in gamma bursting argues against contamination of gamma by spiking activity, and in favor of a more sparse and selective neural code in these layers.

During the WM delay, both alpha/beta bursts and coupling between deep layer alpha/beta and superficial layer oscillations decreased relative to baseline. Alpha/beta oscillations are purported to be an inhibitory rhythm responsible for suppressing behaviorally dominant rules and disregarding distracting stimuli (21–24). Low frequencies in the theta and alpha range modulate high-frequency gamma activity (9, 25–27). We hypothesize that low-frequency coupling between deep and superficial layers may serve a control function, by suppressing access to superficial layers via rhythmic alpha/beta inhibition. The putative control function of deep to superficial layers that we have hypothesized here will be explicitly tested by manipulating WM control in a future study.

WM activity, according to this logic, may be a consequence of deep layer low-frequency modulation of superficial gamma. A decrease in coupling might release inhibition from deep to superficial layers, and allow layer 2/3 spiking and gamma bursting to maintain cue information. There, this information could be stored through recurrent lateral connections that result, on average, in sustained neuronal activity but within a single trial as short-lived bursts of gamma and spiking that reactivate the WM trace (14). This is consistent with PFC anatomy that shows strong and input specific recurrent connectivity within supragranular, but not infragranular, layers (1, 28).

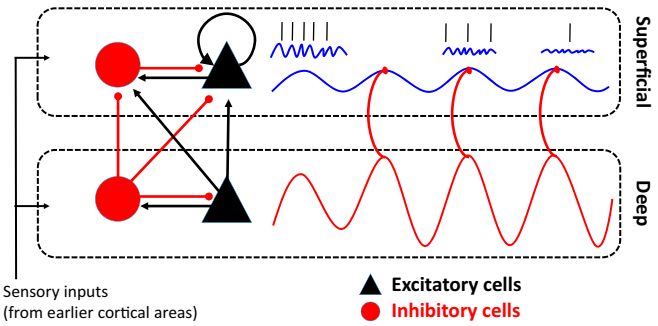


Fig. 6. A model of WM. Denoted by two rectangular, dashed boxes, two cortical compartments, superficial and deep, are made up of densely interconnected pyramidal (black) and inhibitory (red) neurons. Inhibitory connections are line segments with a red, rounded end, and excitatory connections are line segments with a black, arrow end. The looping arrow returning on itself represents the recurrent connectivity found within layer 3 pyramidal cell networks in prefrontal cortex. The sinusoidal red-line in deep layers reflects the predominance of alpha/beta oscillations deep and their driving influence on superficial alpha/beta oscillations (the sinusoidal blue line). Alpha/beta oscillations are coupled with gamma oscillations (blue squiggly lines), and these gamma oscillations organize informative spiking (straight black marks). Over time, moving from left to right in the figure, the deep alpha/beta suppresses both superficial gamma and spiking, which would "clear out" the contents of WM.

In Fig. 6, we summarize this model. We note that both superficial and deep layers are comprised of networks of deeply interconnected excitatory pyramidal (black) neurons and inhibitory (red) interneurons. Circuits in both layers are capable of oscillating within the alpha/beta range (the red sine wave below, the blue line above) but the drive is directional. Deep layers (as seen in the red arrows) drive superficial layers to resonate within the alpha/beta frequency. These alpha/beta oscillations are coupled with superficial layer gamma oscillations. Strong deep to superficial layer coupling and/or deep-layer alpha/beta suppresses gamma-related activity. These dynamics have been previously observed in visual cortex and studied in computational modeling work (29). With a few modifications, this circuit could also implement WM. During the memory delay, we propose that this default suppression of gamma band activity is released, and as a result, the recurrent excitation of layer 2/3 neurons (as indicated by the loop arrow) is allowed to persist. This recurrent excitation generates gamma activity as well as the dominance of a particular ensemble (i.e., one encoding the cue information; ref. 18). We also note that middle and deep layers of PFC are reciprocally connected with the mediodorsal nucleus of the thalamus, with layer 4 receiving thalamic input and layer 5/6 sending output to the thalamus (30). Delay period spiking activity is prominent in MD thalamus (31), and beta band coherence has been reported between PFC and thalamus during WM maintenance (32). Thus, the modulatory role of alpha/beta activity in the deep layers for WM control might be in part regulated by the thalamocortical loop.

Previously, we linked gamma-band dynamics with feedforward mechanisms (21, 33). In that earlier work, gamma was found to signal sensory stimuli from lower to higher visual cortex (6, 33, 34) and to drive stimulus-driven attention (21). Here we find that gamma dynamics are associated with WM maintenance. It has been hypothesized that each cortical area expands on the processing of the previous area with largely conserved laminar circuitry (17) and dynamics (16). In the visual system, the function of superficial layer cells, with gamma band dynamics, is thought to involve feedforward information transmission (35). In PFC cortex, we find preservation of this feature of the laminar circuit (superficial layer gamma-dominated dynamics). At the highest levels of the cortical hierarchy (e.g., PFC), the function of feedforward connections is undefined (36). We suggest that in the absence of further levels to the hierarchy, these superficial layers take on a new role, namely WM.

Experimental Procedures

We performed multilaminar recordings using linear array U and V probes (Plexon). We recorded spiking and LFP activity in frontal and prefrontal cortices of three macaque monkeys (*Macaca mulatta*) while the animals performed tasks requiring either spatial or object-based information to be held in WM. We performed CSD analysis of the LFPs in response to visual stimulation. The earliest reliable current sink was used as the zero point to align sessions. All surgical and animal care procedures were approved by the Massachusetts Institute of Technology's (MIT) Committee on Animal Care and were conducted in accordance with the guidelines of the National Institute of Health and MIT's Department of Comparative Medicine. A detailed descriptions of the study methodology is provided in *SI Appendix, Experimental Procedures*.

ACKNOWLEDGMENTS. We thank Scott Brincat, Morteza Moazami, and Jefferson Roy for assistance during the surgeries and behavioral training; Nancy Kopell and the Cognitive Rhythms Collaborative for fruitful discussions regarding our statistical analyses; and the MIT veterinary staff and animal caretakers for their excellent support. This work was supported by National Institutes of Mental Health Grant R37MH087027, Office of Naval Research Multidisciplinary University Research Initiatives Grant N00014-16-1-2832, and the MIT Picower Institute Innovation Fund.

1. Goldman-Rakic PS (1996) Regional and cellular fractionation of working memory. *Proc Natl Acad Sci USA* 93:13473–13480.
2. Sawaguchi T, Matsumura M, Kubota K (1990) Catecholaminergic effects on neuronal activity related to a delayed response task in monkey prefrontal cortex. *J Neurophysiol* 63:1385–1400.
3. Markowitz DA, Curtis CE, Pesaran B (2015) Multiple component networks support working memory in prefrontal cortex. *Proc Natl Acad Sci USA* 112:11084–11089.
4. Bollimunta A, Mo J, Schroeder CE, Ding M (2011) Neuronal mechanisms and attentional modulation of corticothalamic α oscillations. *J Neurosci* 31:4935–4943.
5. Buffalo EA, Fries P, Landman R, Buschman TJ, Desimone R (2011) Laminar differences in gamma and alpha coherence in the ventral stream. *Proc Natl Acad Sci USA* 108: 11262–11267.
6. van Kerkoerle T, et al. (2014) Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex. *Proc Natl Acad Sci USA* 111: 14332–14341.
7. Maier A, Adams GK, Aura C, Leopold DA (2010) Distinct superficial and deep laminar domains of activity in the visual cortex during rest and stimulation. *Front Syst Neurosci* 4:31.
8. Smith MA, Jia X, Zandvakili A, Kohn A (2013) Laminar dependence of neuronal correlations in visual cortex. *J Neurophysiol* 109:940–947.
9. Spaak E, Bonnefond M, Maier A, Leopold DA, Jensen O (2012) Layer-specific entrainment of γ-band neural activity by the α rhythm in monkey visual cortex. *Curr Biol* 22:2313–2318.
10. Godlove DC, Maier A, Woodman GF, Schall JD (2014) Microcircuitry of agranular frontal cortex: Testing the generality of the canonical cortical microcircuit. *J Neurosci* 34:5355–5369.
11. Ninomiya T, Dougherty K, Godlove DC, Schall JD, Maier A (2015) Microcircuitry of agranular frontal cortex: Contrasting laminar connectivity between occipital and frontal areas. *J Neurophysiol* 113:3242–3255.
12. Miller EK, Erickson CA, Desimone R (1996) Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *J Neurosci* 16:5154–5167.
13. Siegel M, Buschman TJ, Miller EK (2015) Cortical information flow during flexible sensorimotor decisions. *Science* 348:1352–1355.
14. Lundqvist M, et al. (2016) Gamma and beta bursts underlie working memory. *Neuron* 90:152–164.
15. Tort ABL, Komorowski R, Eichenbaum H, Kopell N (2010) Measuring phase-amplitude coupling between neuronal oscillations of different frequencies. *J Neurophysiol* 104: 1195–1210.
16. Bastos AM, et al. (2012) Canonical microcircuits for predictive coding. *Neuron* 76: 695–711.
17. Douglas RJ, Martin KA (1991) A functional microcircuit for cat visual cortex. *J Physiol* 440:735–769.
18. Lundqvist M, Herman P, Lansner A (2011) Theta and gamma power increases and alpha/beta power decreases with memory load in an attractor network model. *J Cogn Neurosci* 23:3008–3020.
19. Sakata S, Harris KD (2009) Laminar structure of spontaneous and sensory-evoked population activity in auditory cortex. *Neuron* 64:404–418.
20. Kampa BM, Letzkus JJ, Stuart GJ (2006) Cortical feed-forward networks for binding different streams of sensory information. *Nat Neurosci* 9:1472–1473.
21. Buschman TJ, Miller EK (2007) Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science* 315:1860–1862.
22. Buschman TJ, Denovellis EL, Diogo C, Bullock D, Miller EK (2012) Synchronous oscillatory neural ensembles for rules in the prefrontal cortex. *Neuron* 76:838–846.
23. Jensen O, Mazaheri A (2010) Shaping functional architecture by oscillatory alpha activity: Gating by inhibition. *Front Hum Neurosci* 4:186.
24. Haegens S, Nácher V, Luna R, Romo R, Jensen O (2011) α-Oscillations in the monkey sensorimotor network influence discrimination performance by rhythmical inhibition of neuronal spiking. *Proc Natl Acad Sci USA* 108:19377–19382.
25. Canolty RT, et al. (2006) High gamma power is phase-locked to theta oscillations in human neocortex. *Science* 313:1626–1628.
26. Lakatos P, et al. (2005) An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. *J Neurophysiol* 94:1904–1911.
27. Colgin LL, et al. (2009) Frequency of gamma oscillations routes flow of information in the hippocampus. *Nature* 462:353–357.
28. Luebke JI (2017) Pyramidal neurons are not generalizable building blocks of cortical networks. *Front Neuroanat* 11:11.
29. Mejias JF, Murray JD, Kennedy H, Wang X-J (2016) Feedforward and feedback frequency-dependent interactions in a large-scale laminar network of the primate cortex. *Sci Adv* 2:e1601335.
30. Giguere M, Goldman-Rakic PS (1988) Mediodorsal nucleus: Areal, laminar, and tangential distribution of afferents and efferents in the frontal lobe of rhesus monkeys. *J Comp Neurol* 277:195–213.
31. Watanabe Y, Funahashi S (2004) Neuronal activity throughout the primate mediodorsal nucleus of the thalamus during oculomotor delayed-responses, I: Cue-, delay-, and response-period activity. *J Neurophysiol* 92:1738–1755.
32. Parnaudeau S, et al. (2013) Inhibition of mediodorsal thalamus disrupts thalamofrontal connectivity and cognition. *Neuron* 77:1151–1162.
33. Bastos AM, et al. (2015) Visual areas exert feedforward and feedback influences through distinct frequency channels. *Neuron* 85:390–401.
34. Michalareas G, et al. (2016) Alpha-beta and gamma rhythms subserve feedback and feedforward influences among human visual cortical areas. *Neuron* 89:384–397.
35. Roberts MJ, et al. (2013) Robust gamma coherence between macaque V1 and V2 by dynamic frequency matching. *Neuron* 78:523–536.
36. Markov NT, et al. (2013) The anatomy of hierarchy: Feedforward and feedback pathways in macaque visual cortex. *J Comp Neurol* 522:225–259.

Supporting Information

Supplemental References

1. Saleem KS, Logothetis NK (2007) *A combined MRI and histology atlas of the rhesus monkey brain in stereotaxic coordinates* (Academic Press, London).

2. Mitzdorf U (1985) Current source-density method and application in cat cerebral cortex: investigation of evoked potentials and EEG phenomena. *Physiol Rev* 65(1):37–100.

3. Schroeder CE, Tenke CE, Givre SJ, Arezzo JC, Vaughan HG (1991) Striate cortical contribution to the surface-recorded pattern-reversal VEP in the alert monkey. *Vision Res* 31(7–8):1143–1157.

4. Lakatos P, Chen C-M, O'Connell MN, Mills A, Schroeder CE (2007) Neuronal oscillations and multisensory interaction in primary auditory cortex. *Neuron* 53(2):279–292.

5. Shipp S (2005) The importance of being agranular: a comparative account of visual and motor cortex. *Philos Trans R Soc Lond B Biol Sci* 360(1456):797–814.

6. Zikopoulos B, Barbas H (2007) Circuits formultisensory integration and attentional modulation through the prefrontal cortex and the thalamic reticular nucleus in primates. *Rev Neurosci* 18(6):417–438.

7. Oostenveld R, Fries P, Maris E, Schoffelen J-M (2011) FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput Intell Neurosci* 2011:156869.

8. Dhamala M, Rangarajan G, Ding M (2008) Estimating Granger Causality from Fourier and Wavelet Transforms of Time Series Data. *Phys Rev Lett* 100(1):018701.

9. Wilson GT (1972) The Factorization of Matricial Spectral Densities. *SIAM J Appl Math* 23(4):420–426.

10. Bastos AM, Schoffelen J-M (2015) A Tutorial Review of Functional Connectivity Analysis Methods and Their Interpretational Pitfalls. *Front Syst Neurosci* 9:175.

11. Trongnetrpunya A, et al. (2015) Assessing Granger Causality in Electrophysiological Data: Removing the Adverse Effects of Common Signals via Bipolar Derivations. *Front Syst Neurosci* 9:189.

12. Vinck M, et al. (2015) How to detect the Granger-causal flow direction in the presence of additive noise? *NeuroImage* 108:301–318.

13. Ray S (2015) Challenges in the quantification and interpretation of spike-LFP relationships. *Curr Opin Neurobiol* 31:111–118.

14. Olejnik S, Algina J (2003) Generalized eta and omega squared statistics: measures of effect size for some common research designs. *Psychol Methods* 8(4):434–447.

15. Sherman SM, Guillery RW (2011) Distinct functions for direct and transthalamic corticocortical connections. *J Neurophysiol* 106:1068–1077.

16. van Kerkoerle T, et al. (2014) Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex. *Proc Natl Acad Sci U S A* 111(40):14332–14341.

17. Godlove DC, Maier A, Woodman GF, Schall JD (2014) Microcircuitry of agranular frontal cortex: testing the generality of the canonical cortical microcircuit. *J Neurosci Off J Soc Neurosci* 34(15):5355–5369.

18. Giguere M, Goldman-Rakic PS (1988) Mediodorsal nucleus: areal, laminar, and tangential distribution of afferents and efferents in the frontal lobe of rhesus monkeys. *J Comp Neurol* 277(2):195–213.

19. Tort ABL, Komorowski R, Eichenbaum H, Kopell N (2010) Measuring phase-amplitude coupling between neuronal oscillations of different frequencies. *J Neurophysiol* 104(2):1195–1210.

20. Lundqvist M, et al. (2016) Gamma and Beta Bursts Underlie Working Memory. *Neuron* 90(1):152–164.

21. Lundqvist M, Herman P, Lansner A (2011) Theta and gamma power increases and alpha/beta power decreases with memory load in an attractor network model. *J Cogn Neurosci* 23(10):3008–3020.

22. Best DJ, Roberts DE (1975) Algorithm AS 89: The Upper Tail Probabilities of Spearman's Rho. Appl Stat 24(3):377.

23. Efron B, Tibshirani RJ (1994) An Introduction to the Bootstrap (CRC Press).

24. Manly BFJ (1991) Randomization, Bootstrap and Monte Carlo Methods in Biology (Second Edition) (Chapman and Hall).

Supplemental Experimental Procedures

Tasks

In order to investigate the mechanisms underlying working memory, we trained three monkeys on three different working memory tasks. Task one (Fig 1A) required memory for a sample object over a delay (which was variable at 0.5-1.2s for 20 recordings and fixed at 1.2s for 9 recordings). The other two tasks engaged working memory for a spatial location; however, they differed in the presentation of a masking stimulus during the delay. In task two (Fig 1B), after a location was cued by a red dot, the animal had to maintain information as a visual mask (red dots at all possible cued locations) was presented through a variable delay (2.2 to 2.7s). In the third task, the monkey had to similarly maintain one of four possible cue locations after a short sample epoch (0.295s); however, during the subsequent 1s delay, there was no visual mask, only the fixation point was shown during the delay. The idea in comparing these tasks was to show that independent of both the content of the working memory and the presence or absence of distractors during the delay, the neural patterns were comparable. Those phenomena we observed, therefore, were general hallmarks of working memory, and for the purposes of this paper all of the data was pooled together (except where indicated). Behavioral performance was high for each of the tasks/monkeys (masked delayed saccade/monkey C: 85%, visual search/monkey S: 77%, delayed saccade/monkey P: 99%). Only correct trials are included in the present analysis, except for Fig. S11, which is an analysis of the error trials pooled over the masked delayed saccade and visual search tasks (there were insufficient error trials from the delayed saccade task).

Recordings

We acutely inserted between 1 and 3 laminar probes into cortex in every recording session. Our recordings included a large portion of the macaque frontal cortex, spanning dorsal premotor regions to lateral prefrontal cortex. In Fig. 1G, the colored dots represent recording sites from each of the monkeys and their respective frontal areas. In addition to the areas that are depicted, we also performed several recording sessions from deeper midline structures, such as the supplementary motor area (SMA) and the anterior cingulate cortex (ACC).

We identified cortical areas based on location relative to anatomical landmarks and used area labels in Saleem and Logothetis (1). MRIs were taken of each individual monkey with the recording grid in place, filled with water, which created a marker to co-register each possible electrode trajectory with the animal's anatomy. Area 8A was defined as recordings on the gyrus posterior to the principal sulcus, and anterior to the arcuate sulcus. This area corresponds to area 8Ad/v in (1). Dorsal Lateral Prefrontal Cortex (DLPFC) were recordings dorsal to the principal sulcus, anterior to Area 8A, and ventral to the upper branch of the arcuate sulcus. This area corresponds to area 8Ad in (1). Ventral Lateral Prefrontal Cortex (VLPFC) were recordings ventral to the principal sulcus, anterior to Area 8A, and dorsal to the lower branch of the arcuate sulcus. This corresponds to area 8Av and 45A in (1). Area Dorsal Premotor (PMd) was defined as the area on the gyrus anterior to the spur of the arcuate sulcus, posterior to the rostral end of the arcuate sulcus, and dorsal to the arcuate sulcus. This area corresponds to cytoarchitectonically defined area F2 and F7 in (1), and is contains part of area SEF

(Supplementary Eye Fields). The deeper midline structures, folded immediately underneath PMd, encompassed the SMA (Supplementary Motor Area) and ACC (Anterior Cingulate Cortex). The ACC corresponds to area 24c in (1). The SMA corresponds to area F3 in (1). Recordings in ACC/SMA were performed in the same trajectories as those for the PMd recordings, but electrodes were advanced deeper (see Fig. S12). Area 8B recordings were performed anterior to the end of the upper branch of the arcuate sulcus and corresponds to area 8Bd in (1). Details on probe spacing, site number, sampling coverage, and areas sampled per monkey/task, are provided in Supplemental Table 1.

We used linear laminar probes ("U probes" and "V probes") from Plexon (Dallas, TX) with a variety of inter-site spacing (100, 150, or 200um) and contact numbers (16, 24, or 32 contacts per probes), giving a sampling range between 1500 and 4500um (see Supplemental Table 1 for details). Probe geometry (inter-site spacing, channel count or U/V type) had no qualitative impact on the data we report here. Because the contact spacing ranged between 100 to 200um, we used cubic spline interpolation to up-sample the data and organize it into the same depth coordinates. To do so, we up-sampled to 100um spacing (if the spacing was already 100um, no interpolation was performed). Each individual probe was considered a unit of observation for our analyses in Figs. 2-5. For the analysis of thresholded units (Figs. S4-7), the unit of observation was thresholded spikes.

In Fig. 1C-E, MRIs are plotted with sample trajectories superimposed (red lines) in each of the monkeys. These trajectories were approximately perpendicular to the cortical sheet, allowing a relatively unbiased sampling of all cortical layers. All of the data was recorded through Blackrock headstages (Blackrock Cereplex M, Salt Lake City, UT), sampled at 30 kHz, band-passed between 0.3 Hz and 7.5 kHz (1^{st} order Butterworth high-pass and 3^{rd} order Butterworth low-pass), and digitized at a 16-bit, 250 nV/bit. All LFPs were recorded with a low-pass 250 Hz Butterworth filter, sampled at 1 kHz, and AC-coupled. In monkey C, the reference was the headpost, in monkey S, the reference was internal to the metal shaft surrounding the probe itself. In monkey P the reference was a nearby guide tube sitting on the dura. Some U/V probes had noisy channels (average power greater than 2 standard deviations above the mean of all channels, this occurred on less than 5% of all channels), which were removed prior to analysis.

Lowering Procedure

In order to place the contacts of the laminar electrode uniformly through the cortex, spanning from cerebrospinal fluid through the gray matter to the white matter, we used a number of physiologic indicators to guide our electrode placement. First, the presence of a slow 1-2 Hz signal, a heartbeat artifact, was often found as we pierced the pia mater and just as we entered the gray matter. Second, as the first contacts of the electrode entered the gray matter, the magnitude of the local field potential increased, and single units and/or neural hash became apparent, both audibly and visually with online spike thresholding. Once the tip of the electrode transitioned into the gray matter, electrodes were lowered slowly with minimal vibration an additional 1-2mm. At this

point, the electrodes were allowed to settle for about 5 minutes, and then we began a visually evoked potential paradigm.

During this paradigm, we flashed a white screen on for 50ms with 500ms pauses while the animal's eye position was detected on the monitor. We repeated this about 200 times, and then calculated a power profile and Current Source Density (CSD) over the contacts of each of the implanted laminar probes. To compute the CSD, we cut the data into trials at the flash onset, obtained the evoked response to the flash across trials, and computed the evoked responses' second spatial derivative (2). The CSD reflects regions where ionic currents are flowing into neurons (caused by a net depolarization of surrounding neurons), and an early sink (where the CSD goes negative within 100ms of stimulus presentation) in response to visual input has been mapped to layer 4 in visual and auditory cortex (3, 4, 16). Layer 4 in these regions receives the bulk of thalamic or bottom-up sensory inputs. In frontal cortex for areas that are agranular (area PMd), an early sink corresponds to the bottom of layer 3 (17), which receives visual sensory inputs (5). In the case of granular prefrontal cortex (containing a well-defined layer 4: areas DLPFC, VLPFC, and 8A/B) we expect the early sink to correspond to layer 4, which receives thalamic input from mediodorsal nucleus of the thalamus as well as sensory cortical areas (18, 5).

We sought to position this current sink at approximately the middle contact of the probe. In addition to using the CSD for electrode alignment, which was more or less noisy, we also examined whether power in the alpha/beta (10-30 Hz), and gamma bands (30-80 Hz) varied across the probe. This was used to determine whether the probe had entered cortex, which was accompanied by a dramatic increase in LFP power across frequencies. The power variance also confirmed whether we had placed the probe in a cortical gyrus (recordings in cortical sulci had minimal power variance), power variance over contact was minimal. We visualized these power estimates at each contact and normalized the power estimates by the maximal power within both of these bands. We found that the crossings of the normalized powers within the alpha/beta and gamma bands nearly always occurred within one or two sites of the earliest sink. Since these power profiles correlated strongly with the CSD results, we used these power profiles to also guide our decisions on whether or not we lowered the electrode any further. Finally, we were also careful not to penetrate by more than an electrode contact or two into the underlying white matter. The white matter was characterized by the predominance of upward going spikes, and the absence of any CSD (i.e. there was markedly low LFP variability). Thus, all of these physiological markers (heart beat artifact, CSD to screen flashes, increase in power at all frequencies upon leaving CSF and entering cortex, power variance in the alpha/beta and gamma bands, presence of spikes/MUA hash, and white matter signatures of lower LFP variance and upward going spikes) were used to determine probe placement. Once the probe was fully lowered, we allowed it to settle for an hour, and then began a longer flashing sequence of 800-1000 trials for offline CSD analysis.

Importantly, the laminar alignment, e.g., each session's "zero point" (corresponding to layer 4/bottom of layer 3), was determined entirely on the basis of offline analysis. It was

based on the site at which current sinks were detected within a time window of 40-180ms of flash onset. In monkey P, due to noise during pre-task flash trials, CSD sinks for alignment purposes were calculated based on the same time window, but in response to sample onset. Across sessions, aligning depth zero to the first significant CSD sink in response to sample onset or the pre-task flash trials gave essentially identical results: the relative gamma and alpha/beta power profiles with respect to the initial sink (Fig. S1) had a Spearman correlation of 1. CSDs were calculated for each trial, taking a spatial integration of between 350-400 um, to be as consistent as possible given our probe geometries (i.e., using 100um inter-electrode spacing, we integrated every 4^{th} contact, for 200um spacing, every 2^{nd} contact, and for 150um, every 3^{rd} contact). We subtracted the CSD at the pre-flash baseline (50ms to flash onset), and then z-scored the CSD data over trials by dividing the raw CSD values by their standard error. We then assessed which CSD contact first achieved a z-score of less than -4 (negative CSD values correspond to current sinks), and which lasted at such a level for at least 6 ms. We assigned this contact as the first significant sink, and this provided the zero (the middle layer) for each penetration.

Relative to this zero point, the average distance to the CSF outside of gray matter was 0.9mm, estimated based on the total power of the LFP (when power across all frequencies decreased below 30% of the contact with maximal power, similar to (17). We also measured the cortical thickness for each of the 60 penetrations based on each individual monkey's MRI. To visualize the specific electrode tracks, we obtained an MRI of each animal with their respective recording chamber and recording grid in place. Filled with water, the grid lumens could be tracked, and electrode trajectories projected onto the cortex and its folds (Fig. 1D-F, Fig. S12). The mean cortical thickness across the penetrations in this study was 2.4mm (measured manually using each animal's MRI, the corresponding recording trajectory, co-registered, visualized, and measured using Osirix software, Geneva, Switzerland). Thus to span the full average cortical distance in our analyses, laminar profiles were plotted from 0.9mm above the sink (the corresponding depth of the average gray matter / CSF transition) to 1.5mm below the sink (average gray matter / white matter transition). Because our recordings span a number of frontal areas that are agranular (containing no discernable layer 4, PMd), dysgranular (containing only a very thin layer 4: SMA and ACC) and granular (containing a well-defined layer 4: 8A, 8B, DLPFC, VLPFC), we chose not to classify particular contacts to specific layers, as the laminar widths and organization could be slightly different over areas. Instead, we chose a more general classification system that could be applied to all areas, grouping contacts at and above the sink as superficial layers (corresponding to layers 1-3) and contacts below the sink as deep (corresponding to layers 5-6). The distance from the sink was a proxy for laminar position.

Analysis

All analysis were performed with customized MATLAB scripts and with Fieldtrip software (7). Given probe movements across a day, we smoothed across contacts on all analyses. This smoothing was symmetric. For the analyses based on LFP and

analog MUA (Fig. 2-5, Figs. S1-3, S8-12), which could be defined for all contacts, we smoothed 200um above and below the contact of interest. This smoothing parameter equaled the distance between the contacts of the coarsest probe used in this study. For the analysis based on thresholded MUA (Figs. S5-7), due to the much sparser nature of the signal, we increased the smoothing to go 500um above and below the contact of interest.

For Fig. 2, power was calculated based on 700 ms segments of data (200 ms prior visual stimulation to 500 ms post visual stimulation), which were tapered with Hanning windows. After these steps, we applied a fast Fourier transform.

For Fig. 3A, we present the MUA change from baseline from 0.150 s post sample offset to 1 second post sample offset. The delay period in this task (search), varied between 0.5 to 1.2 seconds. We determined the average for each time point in the delay based only on trials in which the delay lasted at least until that time point.

To compute phase amplitude coupling (also known as cross-frequency coupling, used in Fig. 5 and S9-10) during the delay or baseline periods we applied the Hilbert transform on two sets of band-passed filtered data: one, band-passed for lower frequencies (4-22 Hz), and the other band-passed at higher frequencies (50-250 Hz). The modulation index (19) was used to quantify the extent to which the faster frequency power deviated from a flat distribution over different phases (taking 18 non-overlapping equally spaced phase bins). We took the phase of the lower frequency Hilbert transform, and the amplitude of the higher frequency Hilbert transform. When no phase-amplitude coupling exists, average gamma amplitude over the 18 non-overlapping alpha/beta phase bins will be similar. This will results in a low entropy value of gamma amplitude over the phase bins. In contrast, when phase-amplitude coupling exists, gamma amplitude will be more concentrated at some phase bins compared to others, resulting in greater entropy over phase. The modulation index is a measure of the degree of entropy (non-uniformity) that is present in this distribution.

The computation of Granger causality in the frequency domain requires the estimation of two quantities: the spectral transfer matrix ($H(\omega)$), which is frequency dependent, and the covariance of the model's residuals (Σ). The spectral transfer matrix defines how power in one channel is transferred to other channels, at each temporal lag. The model's residuals is not a function of frequency, and defines the amount of variance that is left unexplained by the linear model, $H(\omega)$. Traditionally, $H(\omega)$ is computed in a parametric (model-based) fashion by first fitting an autoregressive model to the data, and then Fourier transforming the model (8). However, it is also possible to compute, $H(\omega)$, and thus Granger causality, directly from the spectral transform of the data. In brief, the following fundamental identity holds: $H(\omega)\Sigma H(\omega)^* = S(\omega)$, with $S(\omega)$ being the cross-spectral density matrix at frequency ω. Starting from the cross-spectral density matrix ($S(\omega)$) it is possible to factorize the cross-spectral density matrix into a noise covariance matrix (Σ) and spectral transfer matrix ($H(\omega)$) by applying spectral matrix factorization (9) —which provides the necessary ingredients for calculating Granger causality. The nonparametric estimation of GC has certain advantages over parametric

approaches in that it does not require the specification of a particular autoregressive model order.

Power and cross-frequency coupling analysis were calculated on unipolar data. Granger causality (GC) was computed through a nonparametric spectral matrix factorization of the Fourier transforms of bipolar data (8) during the delay (Fig. 5A). Bipolar derivation is a recommended pre-step prior to Granger causality analysis, as the presence of a common reference can lead to spurious results (10, 11). However, when bipolar derivations are too close to one another and cortical sources are synchronous, the effect of bipolar derivations is to remove the oscillation of interest. Therefore, bipolar derivations were performed between adjacent probes, always taking the contacts that were aligned to the same depth (relative to the early significant CSD sink at each probe). This was only possible when two probes had been simultaneously lowered to the same cortical area (and were between 2-4 mm apart), which occurred a total of 10 times (6 in area PMd in the masked delayed saccade task, 4 in VLPFC in the search task). To control for possible contributions of signal-to-noise differences in driving GC, we performed time-reversed Granger testing (12). In all cases, time-reversing the signals either reduced the dominant directionality or flipped it (from ascending to descending), confirming that signal-to-noise differences could not explain the ascending dominance of GC.

For the analysis of gamma and alpha/beta bursts, we used wavelets to capture deviations in power over time that could change quickly in both time and frequency, reflecting the non-stationarities of interest. Power was computed at equally-spaced frequency bins between 4 and 150 Hz. For each frequency, we used wavelets with a width of 5 cycles, and estimated power every 5ms. Bursts were detected as epochs when power exceeded a threshold of mean +2 standard deviations for at least three cycles, given the center frequency of that burst. Bursts in the gamma-band were detected between 50 to 150 Hz, reflecting the fact that previous findings have shown this to be the upper bound for oscillatory gamma responses (20). Bursts in the alpha/beta range were detected between 4-22 Hz, chosen according to frequencies which had peak power in the deep layers. To minimize the contribution of spikes to the gamma bursts, we removed bursts that contained significant increases in power across all gamma frequencies (50-150Hz), reflecting the fact that spikes are expected to contribute power to this entire band (13) and that bursts should contain power in a relatively narrow frequency range. To calculate whether the delay-period bursting rate contained WM information, we summed the number of bursts around a sliding window (200ms for gamma and 400ms for alpha/beta), and assessed whether the burst rate reliably distinguished between the different cues using an unbiased measure of information, percent explained variance (omega-squared; (14)). Information was then averaged across the delay period.

For the power-power correlation analysis, we correlated power between alpha/beta and gamma in the delay epoch for all possible combinations of alpha/beta providing- and gamma providing-channels, on a trial-by-trial basis (Fig. 5E). We performed this correlation on sessions where the delay length was fixed (total N=13, 4 sessions from

the delayed saccade task and 9 sessions from the visual search task) and monkeys could predict the timing of the test stimulus/go cue. Previous work has shown that predictability of the delay leads to a strengthening of gamma and weakening of alpha/beta (20).

Analysis, MUA Activity

For the analysis of the analog multi-unit activity (MUA, Fig. 3, S3 and S8) we band-pass filtered the raw, unfiltered, 30kHz sampled data into a wide band between 500-5,000Hz, the power range dominated by spikes. The signal was then low-pass filtered at 250Hz and re-sampled to 1,000 kHz. The advantage of using this signal is that every channel can now be used to estimate a MUA signal, even channels which do not contain isolatable spikes.

Analysis, thresholded spikes

For the analysis of thresholded spikes (Figs. S4-7), we began with the raw, unfiltered, 30 kHz sampled data. We re-referenced each contact's signal to the global mean across all contacts, applied a 6^{th} order 250Hz, high-pass Butterworth filter, and then z-scored each signal by its own mean and standard deviation. We next identified spiking by identifying those time periods when the z-scored signal fell below 5 standard deviations of the mean noise floor. In order to ensure that all of these spikes were appropriately captured in time, we further extracted 10 samples before and 24 samples after threshold crossing from the original non-thresholded, 30 kHz signal. This new data was up-sampled by 2 using cubic splines, and the global minimum of each of these snippets was identified. If any spike was counted twice, because of more than a single threshold crossing within this time interval, they were rejected. Finally, to rule out spurious thresholds due to random noise or unit drift, we included units for further analysis if they maintained a firing rate of at least 2 Hz for at least 25 trials. The entire spike analysis pipeline was fully automated.

Analysis, statistics

To assess whether different laminar data features (delay-period MUA and gamma/alpha-beta information, deep-to-superficial CFC vs. other laminar combinations of CFC, ascending vs. descending Granger causality, power-power correlations in different laminar compartments) were significantly different in superficial vs. deep layers, we performed sign tests over laminar probes. The advantage of using a sign test was that the test had minimal assumptions about the underlying distributions. For these different metrics (for example, delay MUA in superficial vs. deep layers in Figure 3H), we first averaged across all superficial and all deep contacts on each session. If superficial MUA was greater than deep, we converted this to 1, otherwise we converted the value to -1. The input to the sign test was the result of this superficial vs. deep comparison (a vector of 1's and -1's). The number of data points was 60 (total number of laminar probes in the study) for comparing relative power in superficial vs. deep

layers (Fig. 2A), delay period MUA in superficial vs. deep layers (Fig. 3H), for comparing delay vs. baseline bursting and information in superficial vs. deep gamma bursts (Fig. 4), for comparing CFC between the 4 different laminar compartments (5C), and for the delay vs. baseline CFC comparison (Fig. 5D).

Power correlation analysis (Fig. 5E) used 13 data sets due to the requirement of a fixed delay-period. Also, Granger causality analysis (Fig. 5A) used 10 data sets, due to the requirement of bipolar re-referencing across probes. Here, the paired sets of data that were compared (and transformed to 1's and -1's) were average GC in the superficial to deep direction vs. average GC in the deep-to-superficial direction.

To compute correlations between laminar profiles, we first computed the average profile for a particular data feature (relative power, cross-frequency coupling, or delay-period MUA modulation), taking the average across either all laminar profiles, all laminar profiles within a task, or all laminar profiles within an area (as indicated in the text). We then computed Spearman's rank correlation (Rho) to assess, for example, the hypothesis that the grand-average gamma and alpha/beta power profiles (Fig. 2B) were correlated to the grand-average delay period MUA profile (Fig. 3G). The reason we used Spearman rank correlation is because the underlying distribution of the data (power/CFC/MUA profiles) was not normally distributed. The magnitude of Spearman's Rho is computed by calculating the Pearson correlation coefficient after rank ordering the two sets of variables. The p-value of this correlation is determined using a non-parametric test that does not make any assumptions about the underlying distribution. The non-parametric test proceeds as follows: a reference distribution is approximated by randomly permuting the laminar order of the data. This is equivalent to destroying the laminar dependence of the data. This realizes the null hypothesis that relative power/CFC/MUA values do not depend on laminar depth. A large-sample approximation of this permutation distribution is then used to compute the p-values we have reported (22). We obtained the same p-values when we manually created the permutation distributions using 10,000 permutations (but with this number of permutations could only resolve p values as low as 1E-4). We report p-values based on this large-sample approximation (22). All correlations were performed on un-smoothed data.

To assess the hypothesis that a given area's gamma/alpha-beta/CFC laminar profile was significantly correlated to the average of the remaining areas (Supplementary Figure 2 and 9), we used bootstrap resampling. This was necessitated by the fact that we wanted to correlate each area's average profile to an unbiased representation of the average of the remaining areas. To calculate an average of the remaining areas that was not biased by differences in sample size between areas, we resampled 10 laminar profiles from each area with replacement, and then took the average. This procedure was repeated 1,000 times to calculate a bootstrap distribution of Spearman rank correlation values between the average of each area and the (unbiased) average of the remaining areas. The reason we resampled with replacement was to approximate as closely as possible the classic bootstrap procedure (23). The reason we resampled 10 laminar profiles for each area was because the average number of laminar profiles per area was 10 in our study. Note that some areas were sampled less than 10 times, and

so the resampling for these areas will re-use the same laminar profile many times (in order to equally weigh the areas with less sampling in the average). Changing the number of area laminar profile resamples per bootstrap to 5 or 2 did not qualitatively change the results. We then used the formula from Efron and Tibshirani (23, equation 2.3) to calculate the standard error of the mean of this bootstrap distribution, and computed the 95% confidence intervals by multiplying this estimate of the standard error by 1.96. P-values for the bootstrap average correlation coefficient were determined by shifting the bootstrap distribution to zero by subtracting its mean. The p-value then corresponds to the rank of the average correlation coefficient relative to this zero-centered distribution divided by the number of bootstrap re-samples (24, equation 3.10).

Supplemental Results

Analysis of isolated units and spike rates confirm delay-period activity in superficial layers

Fig. S5A shows the number of times a laminar depth was sampled (in black), the number of units recorded at each depth (in green), and the number of units that were modulated (in blue), summed across sessions. Modulation was determined by testing whether each unit's firing rate during the delay was significantly different from its baseline firing rate by a two-tailed t-test at $p<0.01$. Fig. S5B shows the proportion of units with delay activity (i.e., units modulated/units recorded). This was significantly different between superficial vs deep layers (Fig. S5B inset, 53% vs. 34%, Chi-squared proportion test, p=3E-4). The proportion of modulated units by layer (Fig. S5B) positively correlated with the gamma LFP laminar profile and negatively correlated with the alpha/beta LFP profile (Spearman rank correlation, gamma, rho = 0.77, p<2E-4, alpha/beta, Rho = -0.67, p=0.002). Spiking during the delay period carried significant information about the sample (measured using Percent Explained Variance, PEV, see Methods). The peak spike PEV value was observed in superficial layers at depth -400um, the same depth at which gamma power peaked. The more alpha/beta a layer contained, the less information it contained about the sample (Spearman rank correlation, Rho = -0.48, p=0.02, Fig. S6). The greater proportion of modulated units in superficial layers was not the result of a poor signal to noise ratio or lack of units in deep layers. In fact, baseline firing rates were higher in the deep layers (Fig. S5C), as was the unit yield (Fig. S5D). A similar delay-period profile was seen when we aligned depth based on the transition between cerebrospinal fluid and the gray matter (Fig. S7).

Sample onset and saccade processing activate spiking in middle and deep layers

Anatomical and physiological studies of the cortical microcircuit have shown that feedforward inputs arrive in layer 4, and the output of cortical computation to motor structures is signaled via deep layers, especially layer 5 (15). Therefore, we investigated whether the sample processing and saccade execution stages of the task

would activate these other layers. For the granular cortical areas (VLPFC, DLPFC, 8A, sampled in monkeys S and P), the sensory MUA response at sample onset was strongest at depths near zero (depth -200 um for monkey S, 100 um for monkey P, Fig. S8 A/E). In addition, all areas/monkeys showed robust deep layer activation just preceding the saccade (Fig. S8 B/D/F/H). The ratio of MUA activity in deep layers to superficial layers was significantly larger in the 50 ms prior to the saccade compared to the delay period (sign test, $p<0.01$, Fig. S8H/I). This indicates that the relative balance in cortex switches from superficial-layer-dominant during the delay to driven by both deep and superficial layers during the saccade period.

Recurring dynamics across all six cortical areas

To test the hypothesis that certain aspects of the laminar dynamics are shared across areas, we calculated each area's average gamma and alpha/beta relative power profile and each area's cross-frequency coupling (CFC) profile between deep layers' alpha/beta phase to all other layers' gamma amplitude. Each area's average profile was taken across all available data from that area, and is shown as the bold lines in the upper subpanels of Fig. S2 and S9. We reasoned that if there are shared dynamics, then the average of each area should be correlated to the average of the remaining areas. To test this, we calculated Spearman rank correlation between each area's average power/CFC profile and the average of the remaining areas (excluding the current area, so as to perform correlation on independent data, and equally weighing each area when computing the other-area average, see Analysis, statistics). We assessed significance of the correlation using a bootstrap resampling procedure (see Analysis, statistics). In SI Figures 2 and 9, we plotted these bootstrap distributions and their 95% confidence intervals. In SI Figure 2, the gamma and alpha/beta relative power profiles of each area are significantly correlated to the unbiased other-area-average (Bootstrap test, $p<0.01$) in all cases except one (alpha/beta profile of area 8B). In SI Figure 9, the CFC profile of each area is significantly correlated to the unbiased other-area-average in all cases (Bootstrap test, $p<0.01$).

Finally, we determined the significant of observing the peak delay-period MUA modulation in superficial areas for all areas under the null hypothesis that the peak could occur in either superficial or deep layers with a sign test over the six areas ($p=0.03$, SI Figure 3).

Spiking and bursting are modulated on error trials

To add evidence for the functional relevance of our measures of MUA modulation and LFP bursts, we analyzed the error trials. Compared to error trials, correct trials had elevated MUA in deep layers (sign test, $p<0.05$, Bonferroni corrected, Fig. S11A). Correct trials were associated with less LFP gamma bursting during the delay (sign test, $p<0.05$) but more alpha/beta bursting (sign test, $p<0.05$) during the baseline (Fig. S11B). These results indicate that the strength of delay period MUA modulation and the strength of alpha/beta and gamma bursting are relevant for correct WM performance.

These results suggest that both superficial and deep layers are important for correct WM performance, but in distinct ways. As we show in Fig. 4B, gamma bursting during the delay in superficial layers is informative about the sample (in correct trials). Deep layer alpha/beta rhythms modulate this gamma (Fig. 5B). This deep to superficial modulation has an inhibitory effect on superficial gamma (Figure 5E). Thus, errors could come about by a combination of mechanisms: less (inhibition-related) alpha/beta bursting during the baseline would make the default state of the cortex overly excited with internal information, and less receptive to the sample (on error trials). This is born out in the delay, where we observe an excessive amount of gamma bursting in error trials, and a relative lack of deep layer MUA. Thus, we interpret error trials as a failure of control-related deep layer activity. This results in abberant modulation of WM information-related gamma.

Supplemental Discussion

Sustained activity and Working Memory

In all three tasks that we studied, we observed changes in MUA activity during the delay vs. baseline. Two of the tasks (involving spatial WM) also involved maintaining a motor plan, while the object WM task did not. In all cases, we observed modulation in average MUA activity in superficial layers. This is evidence that the MUA modulation is not a trivial consequence of oculomotor signals (which are not informative in the delay period of the object based task). This suggests that neuronal activity is regulated during WM delays, although not necessarily in a sustained pattern of firing. Indeed, recent models suggest that WM information is encoded in a pattern of synaptic strength, and is periodically refreshed during the occurrence of gamma oscillations (20, 21). Thus, elevated spiking is not necessary throughout the trial. Our finding of a strong correlation between layers with gamma and layers with delay activity supports this model, and suggests that this mechanism occurs predominately in superficial layers.

Supplemental Figures Captions

Supplemental Figure 1. ***A.*** The grand average current source density analyses calculated with respect to sample onset (during working memory task performance) or ***B*** after flash onset (during pre-task mapping) and after alignment to the first significant sink on each electrode. The white contour lines reflects significance at $p < 2.4E-5$ (Bonferroni correction). ***C.*** Power asymmetry in the gamma and alpha/beta bands calculated based on data epochs from 200ms pre-sample onset to 500ms post-sample onset (during task performance). ***D.*** Power asymmetry in the gamma and alpha/beta bands calculated based on data epochs from 200ms pre-flash onset to 500ms post-flash onset (during pre-task mapping). Both gamma and alpha/beta profiles in ***C*** and ***D*** have a Spearman correlation of 1 ($P<5E-7$).

Supplemental Figure 2. (Upper subpanels) Relative power profiles in the gamma (red, thick lines) and alpha/beta (blue, thick lines) bands across areas, in order (left to right

and top to bottom) of areas with more to less sampling (N, which is the number of laminar probes per area). The less saturated and thinner red/blue lines are the bootstrapped re-sampled and bias-free average of the remaining areas (other-area averages were corrected for the differences in N between areas). The solid and thin lines for each area were correlated, and the average (over bootstrap resamples) resulting of Spearman rank correlation (Rho) and their p-values are plotted above each upper subplot. (Lower subpanels) The distribution of bootstrap resampled Spearman's Rho between each area's average power profile and the unbiased-other-area average are plotted separately for alpha/beta (blue) and gamma (red). The 95% confidence interval for these distributions are shown in the thick red/blue lines above the distribution. For the areas with sparse N (bottom row), the y-axis has been adjusted to show laminar depths which were consistently recorded in that area.

Supplemental Figure 3. The average delay-period MUA modulation per cortical depth across all areas we sampled, in order (left to right and top to bottom) of areas with more to less sampling. For the areas with sparse sampling (bottom row), the y-axis has been adjusted to show laminar depths which were consistently recorded in that area.

Supplemental Figure 4. A representative session's peri-stimulus time histogram (PSTH) depicting average firing rate +/- 1SEM in dotted lines across the trial for the masked delayed saccade task. The task progression (fixation, sample onset, mask onset, delay) is shown in the upper PSTH. Four units are shown (2 in superficial layers, 2 in deep layers) along with their depth on the laminar probe, indicated by the arrows. Above each unit's PSTH is the unit's unique identifying code and the number of trials for which the respective unit was observed. To the right of each PSTH, a random sample of 1,000 individual spike waveforms are shown for each respective unit (the total number of spikes for which the unit was observed is depicted above each waveform).

Supplemental Figure 5. ***A.*** Number of times a depth was sampled (in red), number of units detected (in green) and the absolute number of units significantly modulated (in blue) during the delay (t-test, $p<.01$) across laminar depths. ***B.*** The proportion of units that were significantly modulated during the delay across depths. (inset) The proportion modulated in superficial vs. deep layers. Error bars $\pm$ 1 SEM. ***C.*** The mean baseline firing rates across laminar depths. ***D.*** The proportion of electrodes with detectable units.

Supplemental Figure 6. Power asymmetry in the gamma (red) and alpha/beta (blue) bands over laminar depth (same as in Figure 2B). In black, the information (calculated with PEV) in delay-period spiking about the sample over laminar depth is plotted.

Supplemental Figure 7. The proportion of units that were significantly modulated during the delay for each laminar depth, aligned to the transition zone between cerebrospinal fluid (CSF) and the gray matter (depth 0). This transition zone was estimated as the contact where a sudden drop in power of 70% or more across all LFP frequencies occured. This was only possible in recording sessions where at least one channel remained outside the cortex in CSF (N=24 sessions). For this subset of

sessions, the proportion of units with delay activity showed a similar profile as the full dataset, peaking in more superficial layers, and decreasing with depth.

Supplemental Figure 8. ***A/C/E (left).*** Average MUA modulation with respect to fixation baseline, time-locked to sample onset and plotted over time and laminar depth, for each respective task. ***A/C/E (right).*** Same as ***A/C/E (left)*****,** but average over the time window indicated above the graphs, from 50-100ms post sample onset (in black) and for the first 1 second of the delay period (in blue). ***B/D/F (left).*** Average MUA modulation with respect to fixation baseline, time-locked to saccade execution and plotted over time and laminar depth, for each respective task. ***B/D/F (right).*** Same as ***B/D/F (left)*****,** but average over the time window indicated above the graphs as the saccade window from 0-50ms pre-saccade (in red) and the end of the delay period (in blue).

Supplemental Figure 9. (Upper subpanels) The average delay-period Cross-Frequency Coupling of each area (thick blue lines). The Cross-Frequency Coupling was quantified using the modulation index (see Methods). The modulation index profile was calculated by averaging across deep layers' (depths greater than +100 um alpha/beta phase) alpha/beta phase, and plotting the resulting CFC modulation of the gamma amplitude of all other layers. Areas are plotted in order (left to right and top to bottom) with more to less sampling (N, which is the number of laminar probes per area). The thinner gray lines are the bootstrapped re-sampled and bias-free average of the remaining areas (other-area averages were corrected for the differences in N between areas). The solid and thin lines for each area were correlated, and the average (over bootstrap resamples) resulting of Spearman rank correlation (Rho) and its p-value are plotted above each upper subplot. (Lower subpanels) The distribution of bootstrap resampled Spearman's Rho between each area's average CFC profile and the unbiased-other-area average is plotted. The 95% confidence interval for these distributions are shown in the thick black lines above the distribution. For the areas with sparse N (bottom row), the y-axis has been adjusted to show laminar depths which were consistently recorded in that area.

Supplemental Figure 10. ***A.*** Cross Frequency Coupling (CFC) between the phase of alpha/beta oscillations and the amplitude of gamma oscillations during the baseline period. Plotted across both axes is the CFC between specific cortical depths. ***B.*** The mean CFC across all four possible conditions during the baseline period: superficial phase to superficial amplitude (left), deep phase to deep amplitude (middle-left), superficial phase to deep amplitude (middle-right), and deep phase to superficial amplitude (right).

Supplemental Figure 11. ***A.*** The difference between the delay-period MUA modulation of correct trials minus error trials per laminar depth, averaged across sessions with at least 20 error trials (N=56). Error bars +/- 1 SEM for illustration. The asterisks denotes the depths at which the differences were significant at $P<0.05$ (sign test across sessions, red, Bonferonni corrected for multiple comparisons, black uncorrected). ***B.*** The average bursting rate across layers and sessions in the alpha/beta (left) and gamma (right) bands, during the baseline fixation period (upper graphs) and during the

delay period (lower graphs). Error bars +/- 1 SEM for illustration. Asterisks denotes significant differences at P<0.05 (sign test over sessions).

Supplemental Figure 12. *A*. Laminar profile of gamma and alpha/beta LFP over recording depth in area SMA. In this case, layers are inverted with respect to channel depth: deep cortical layers (L5/6) are in more superficial channel depths, and superficial cortical layers (L2/3) are in deeper channel depths ***B.*** Laminar profile of gamma and alpha/beta LFP over layers in ACC: the power profile re-normalizes, with superficial gamma in superficial channel depths, and deep alpha/beta in deep channel depths. ***C*.** An MRI slice of frontal cortex, aligned to the recording grid (white lines). The red lines and arrow correspond to a continuation of the white lines, and therefore indicate approximate recording trajectories. Recording trajectories were chosen to be as perpendicular as possible to cortex. The left most trajectory could also be used to access the deeper cortical areas (SMA and ACC). Area labels are approximate and based on (1) - coronal slice +28mm rostral to Ear Bar Zero. The MRI is in the same plane as the recording grid, therefore the sulcus to the right of the trajectories corresponds to the genu of the arcuate sulcus. ***D.*** The recording sites used in this study relative to anatomical landmarks. The dotted black line corresponds to the sample trajectories shown in ***C***. ***E.*** Laminar profile of gamma and alpha/beta LFP over recording depth in area PMd for a representative session.

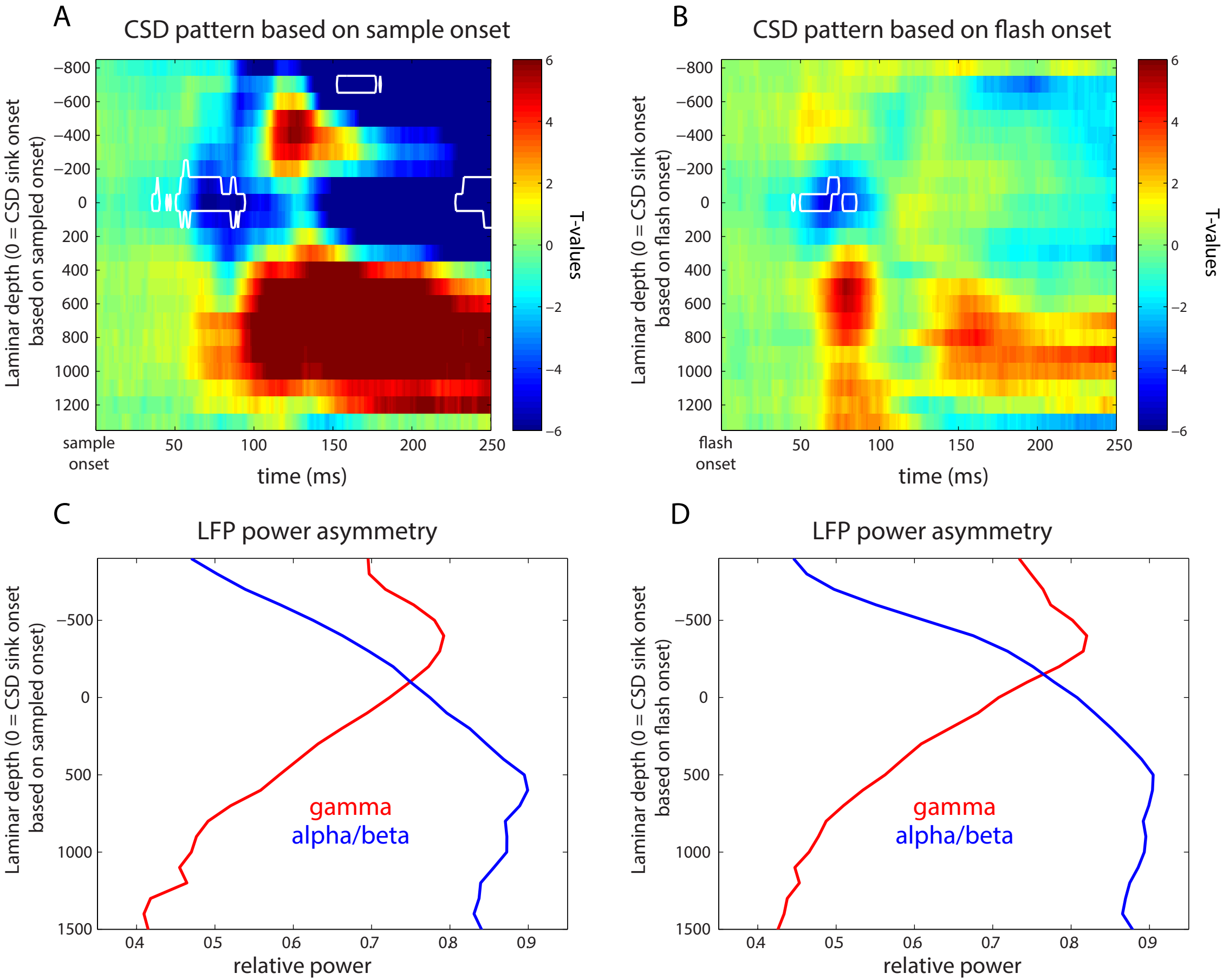
Supplemental Tables

Supplemental Table 1. A table listing information about each of the 60 recordings we performed. From left to right, we list the number of channels on the U/V probe, the inter-electrode distance, the total coverage of the probe, the cortical area sampled, the monkey recorded, and the task performed.

Supplemental Table 1

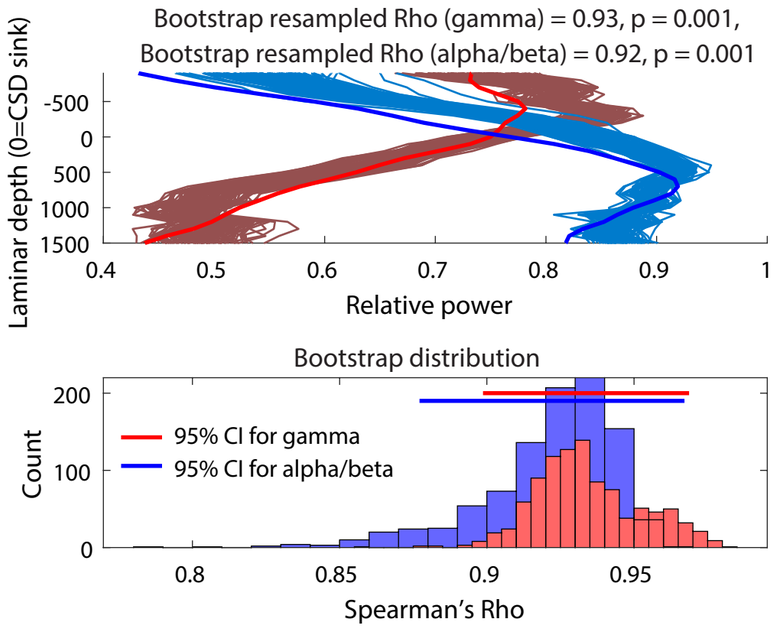
Num channels on probe	Spacing (um)	Coverage (um)	Area	Monkey	Task
16	200	3000	PMd	C	Masked Delayed Saccade
16	150	2250	PMd	C	Masked Delayed Saccade
16	150	2250	PMd	C	Masked Delayed Saccade
16	200	3000	PMd	C	Masked Delayed Saccade
16	200	3000	PMd	C	Masked Delayed Saccade
16	150	2250	PMd	C	Masked Delayed Saccade
16	200	3000	PMd	C	Masked Delayed Saccade
16	150	2250	PMd	C	Masked Delayed Saccade
16	200	3000	PMd	C	Masked Delayed Saccade
16	150	2250	PMd	C	Masked Delayed Saccade
16	200	3000	PMd	C	Masked Delayed Saccade
16	200	3000	PMd	C	Masked Delayed Saccade
16	150	2250	PMd	C	Masked Delayed Saccade
16	200	3000	ACC	C	Masked Delayed Saccade
16	150	2250	ACC	C	Masked Delayed Saccade
16	200	3000	PMd	C	Masked Delayed Saccade
16	150	2250	SMA	C	Masked Delayed Saccade
16	200	3000	PMd	C	Masked Delayed Saccade
16	200	3000	8B	C	Masked Delayed Saccade
24	150	3450	8B	C	Masked Delayed Saccade
32	100	3100	PMd	C	Masked Delayed Saccade
32	100	3100	PMd	C	Masked Delayed Saccade
32	100	3100	SMA	C	Masked Delayed Saccade
32	100	3100	PMd	C	Masked Delayed Saccade
16	200	3000	8B	C	Masked Delayed Saccade
16	150	2250	PMd	C	Masked Delayed Saccade
16	150	2250	PMd	C	Masked Delayed Saccade
16	200	3000	VLPFC	S	Search
16	200	3000	VLPFC	S	Search
16	200	3000	VLPFC	S	Search
16	200	3000	8A	S	Search
16	200	3000	VLPFC	S	Search
16	200	3000	8A	S	Search
16	200	3000	8A	S	Search
16	200	3000	VLPFC	S	Search
16	200	3000	8A	S	Search
16	200	3000	VLPFC	S	Search
16	200	3000	VLPFC	S	Search

16	200	3000	VLPFC	S	Search
16	200	3000	VLPFC	S	Search
16	200	3000	VLPFC	S	Search
16	200	3000	VLPFC	S	Search
16	200	3000	DLPFC	S	Search
16	200	3000	DLPFC	S	Search
16	200	3000	8A	S	Search
16	200	3000	8A	S	Search
16	200	3000	8A	S	Search
16	200	3000	8A	S	Search
16	200	3000	VLPFC	S	Search
16	200	3000	VLPFC	S	Search
16	200	3000	VLPFC	S	Search
16	200	3000	VLPFC	S	Search
16	200	3000	VLPFC	S	Search
16	200	3000	VLPFC	S	Search
16	200	3000	VLPFC	S	Search
16	200	3000	VLPFC	S	Search
32	100	3100	VLPFC	P	Delayed Saccade
32	100	3100	VLPFC	P	Delayed Saccade
32	100	3100	VLPFC	P	Delayed Saccade
32	100	3100	VLPFC	P	Delayed Saccade

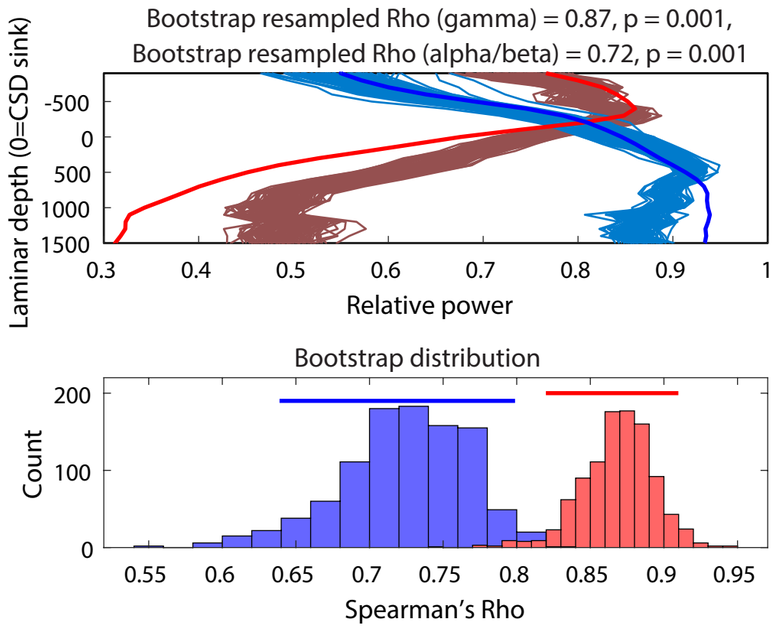


Supplemental Figure 1

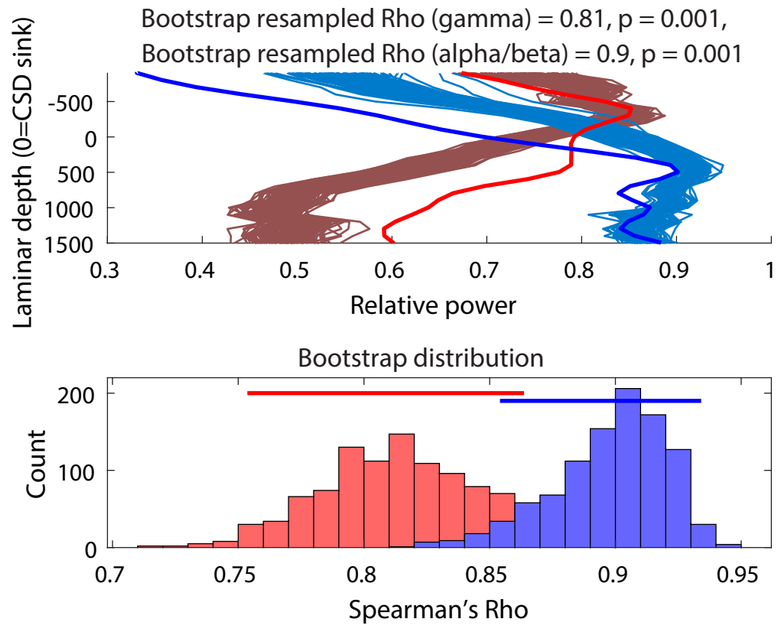
VLPFC (N=23)



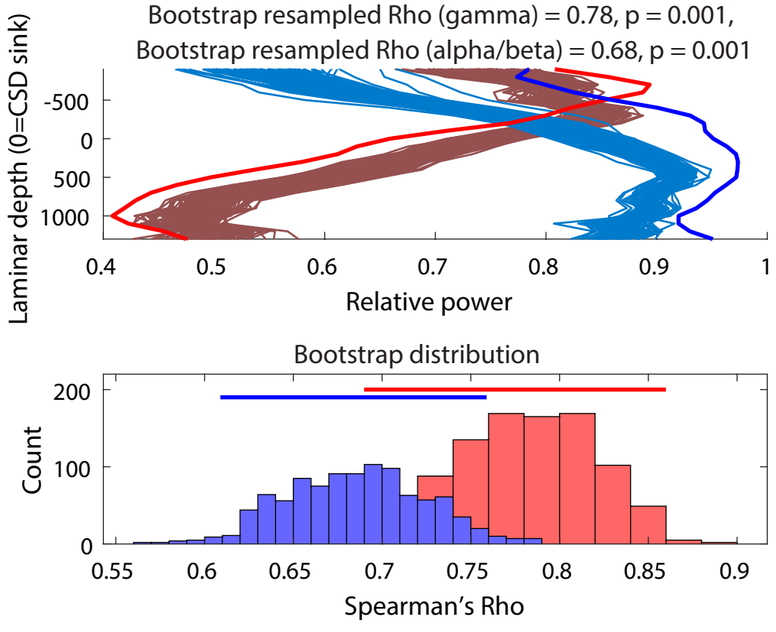
PMd (N=20)



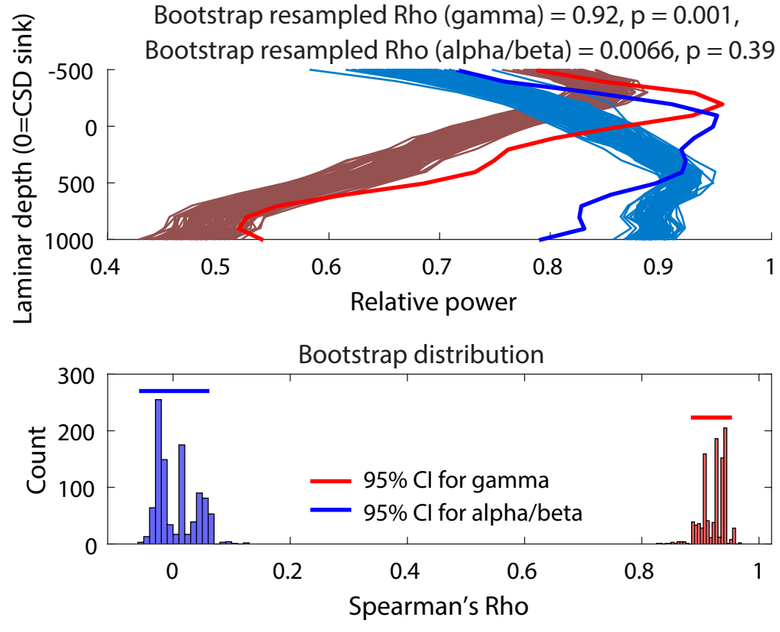
8A (N=8)



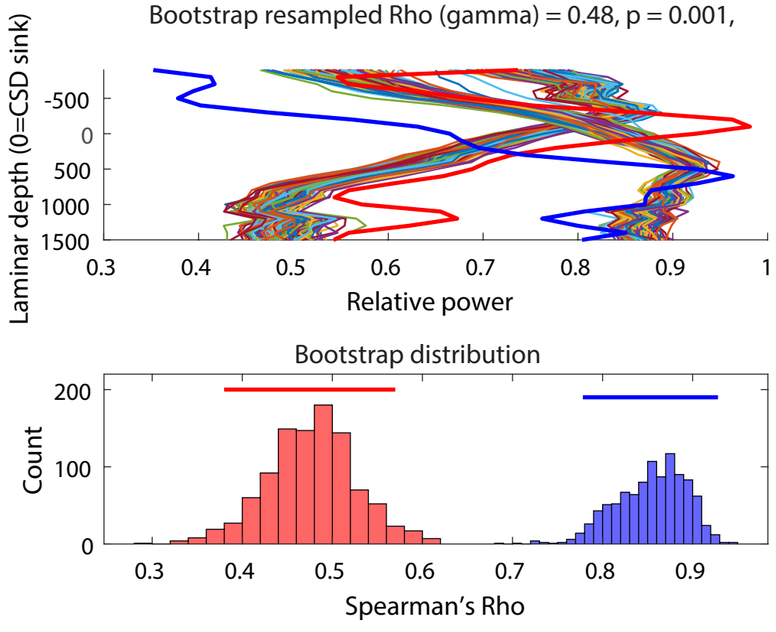
ACC/SMA (N=4)



8B (N=3)



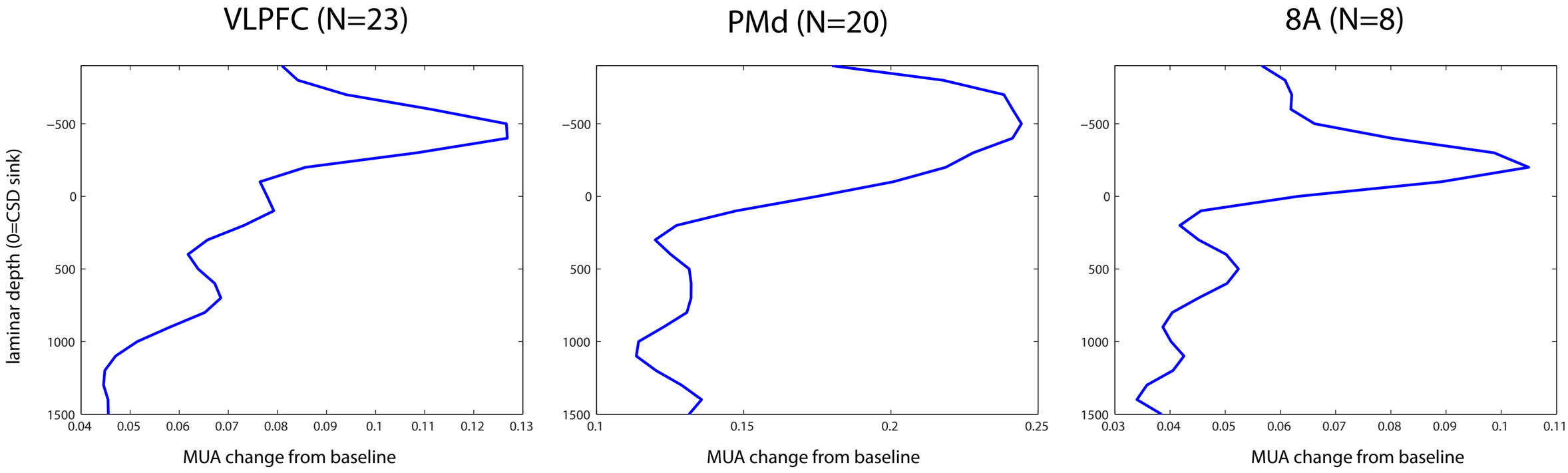
DLPFC (N=2)



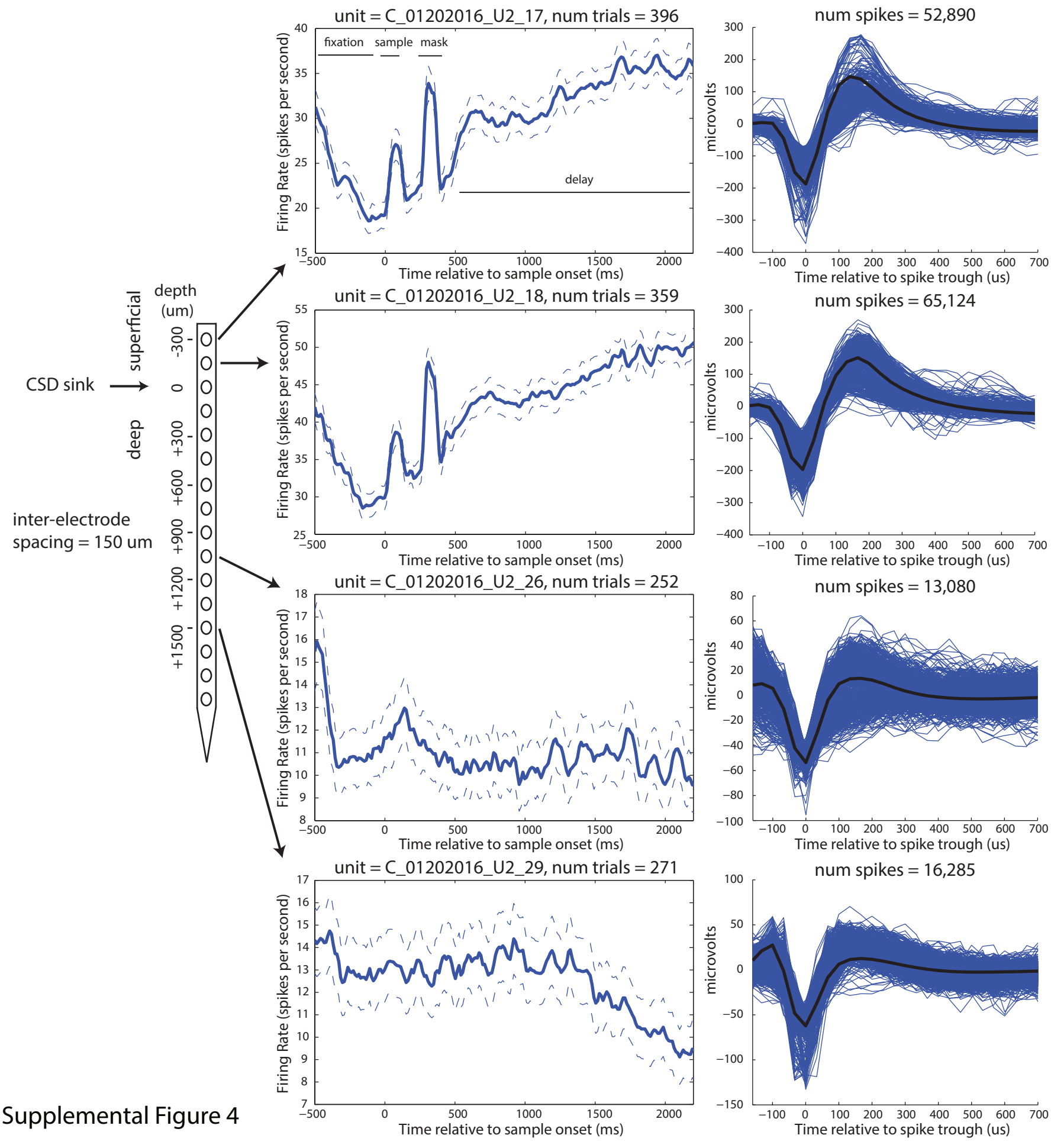
Supplemental Figure 2

Legend (upper subplots)

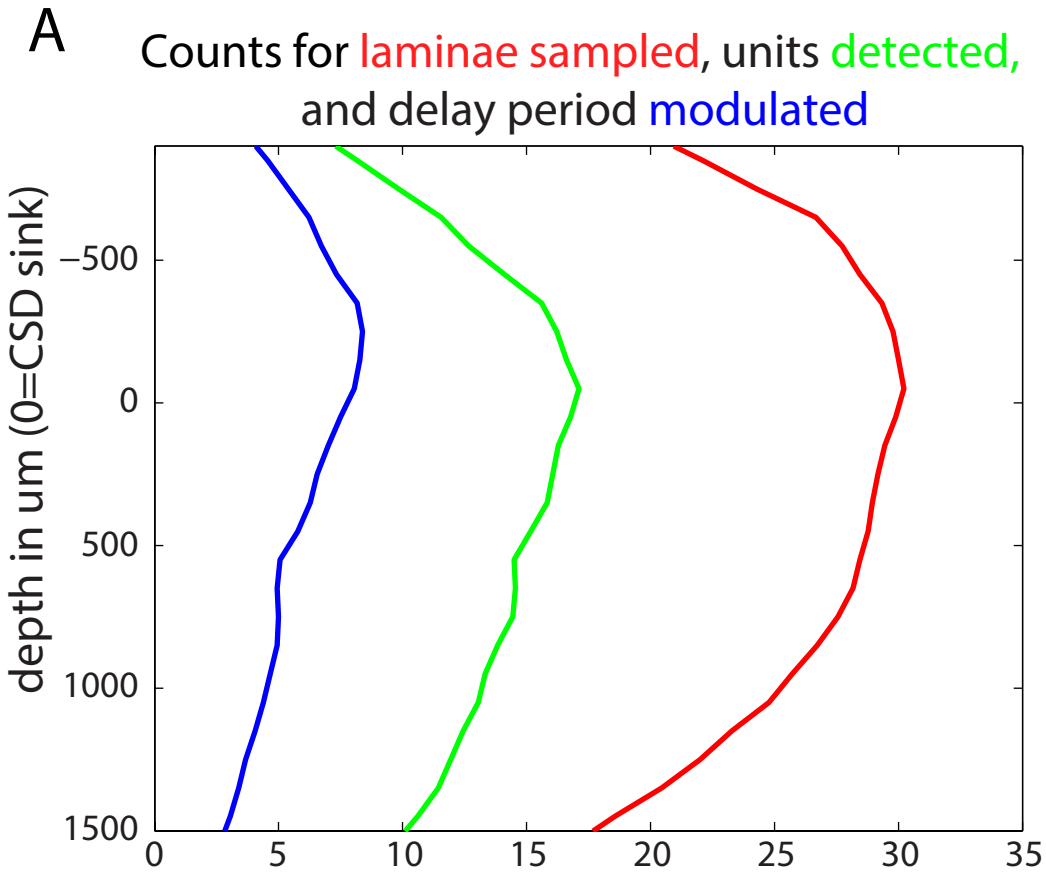
- Gamma power (average of each area)
- Gamma power (resampled average of all other areas)
- Alpha/beta power (average of each area)
- Alpha/beta power (resampled average of all other areas)

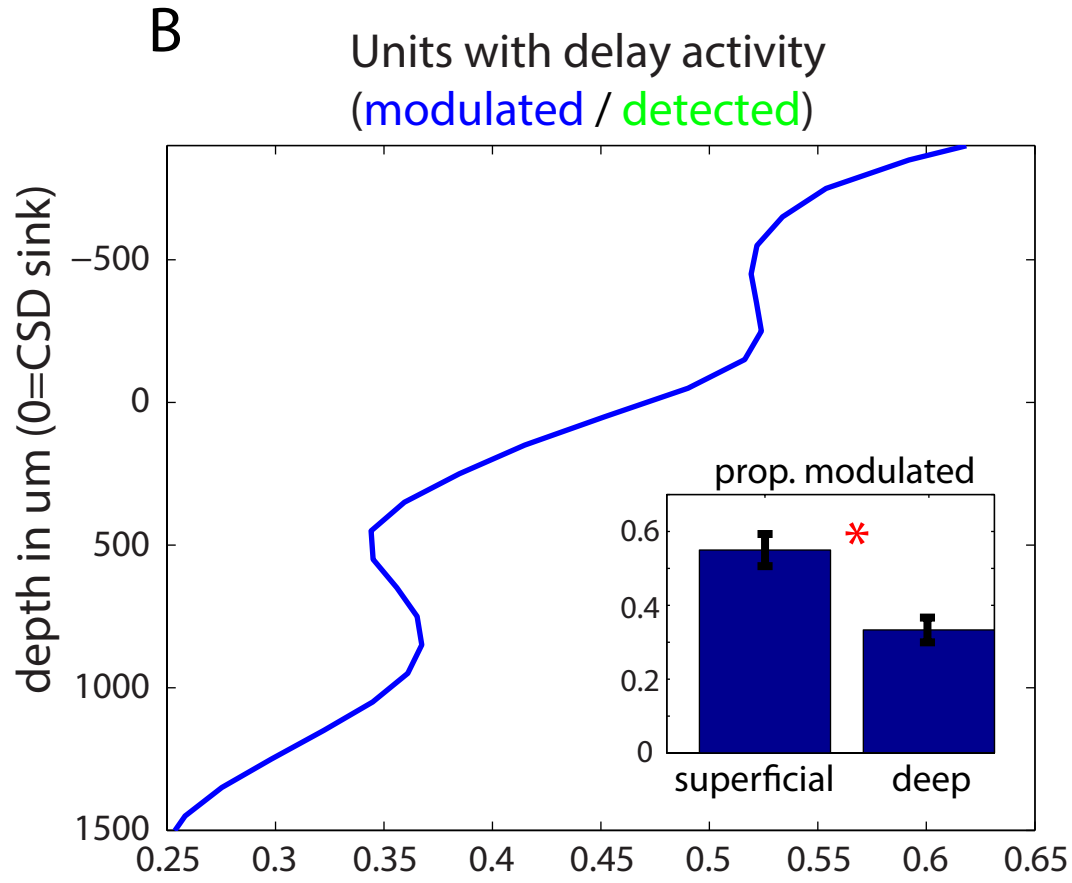


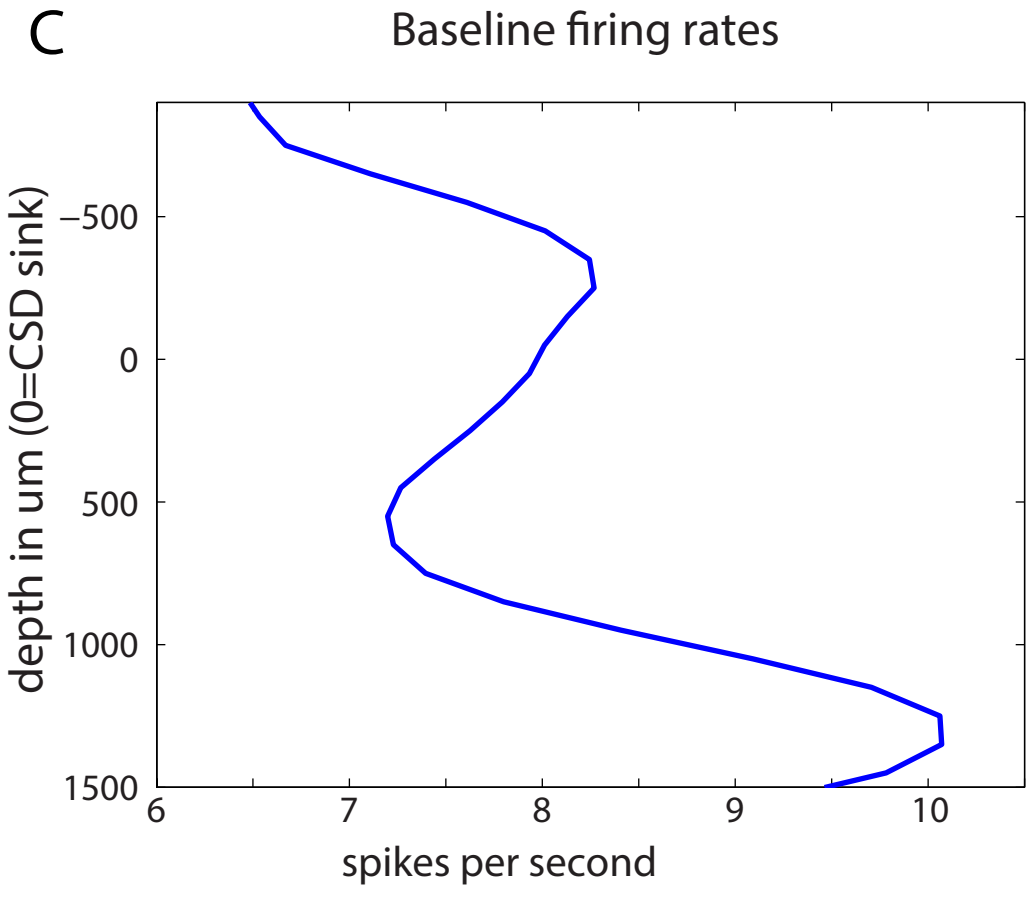
Supplemental Figure 3

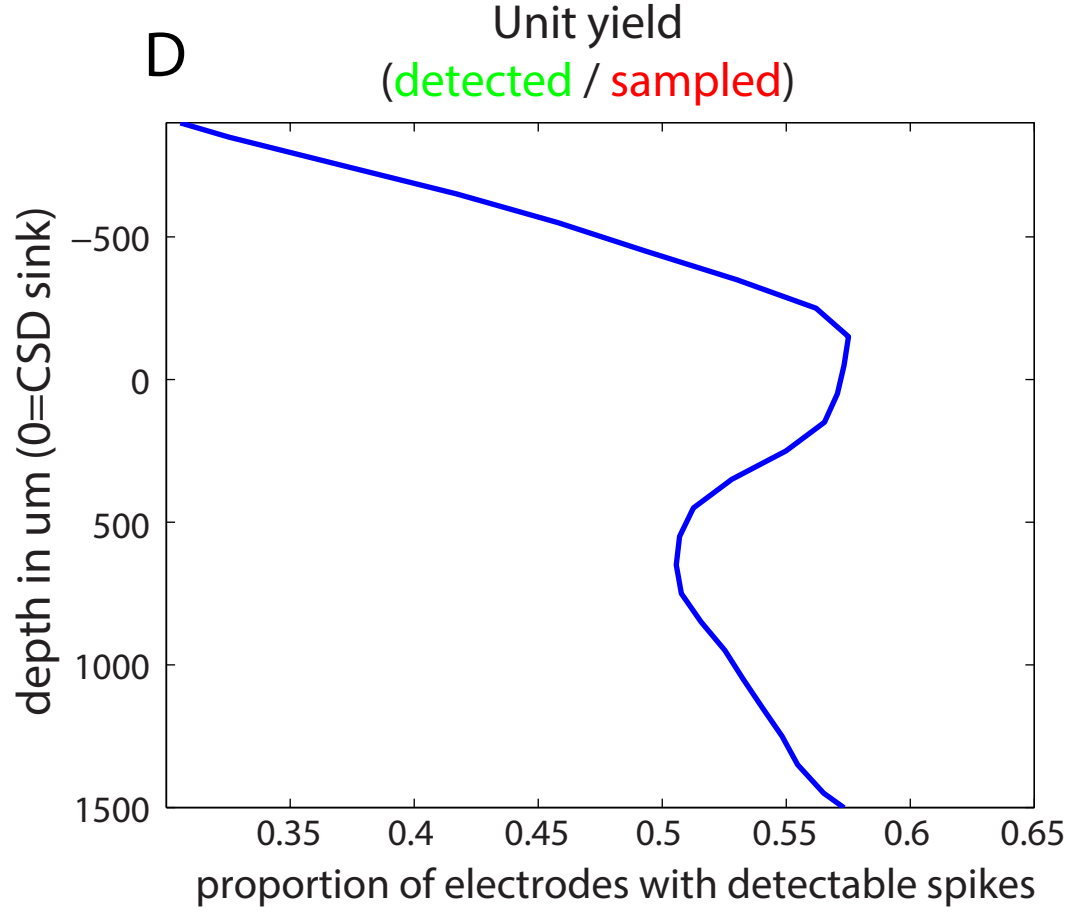


Supplemental Figure 4

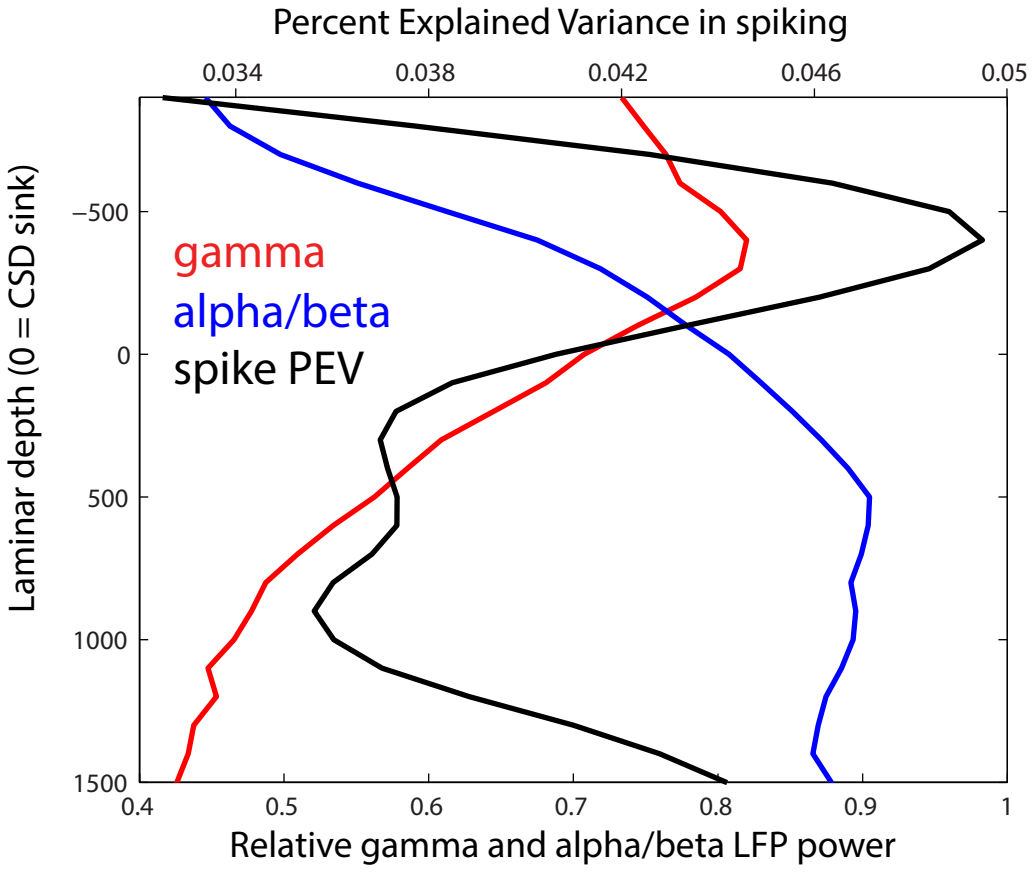




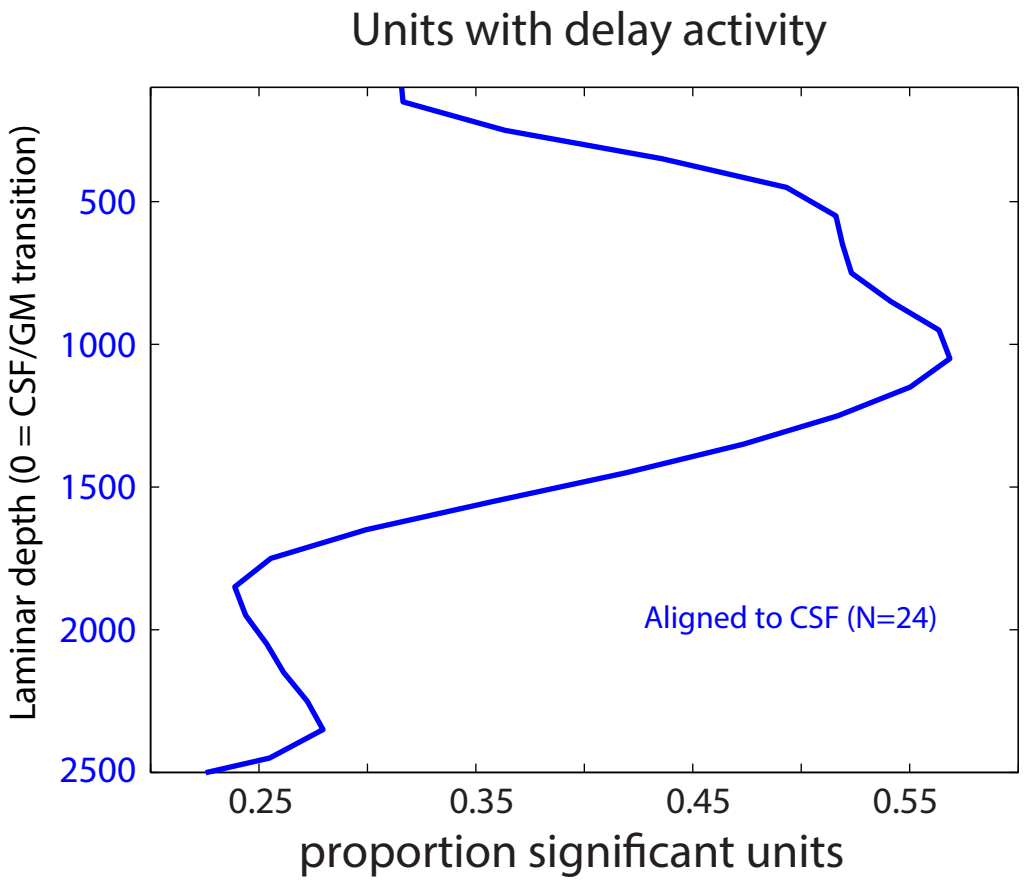




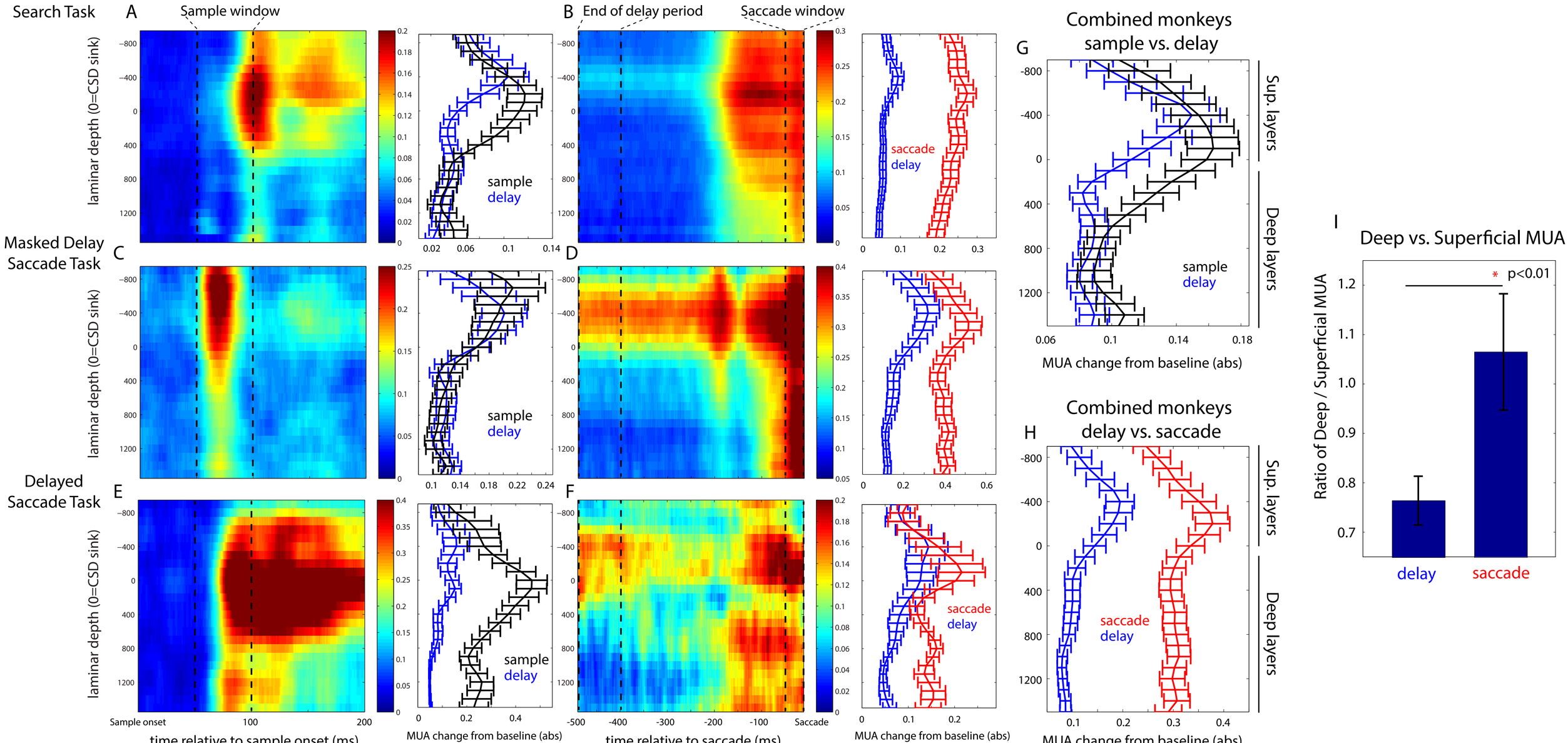
Supplemental Figure 5



Supplemental Figure 6

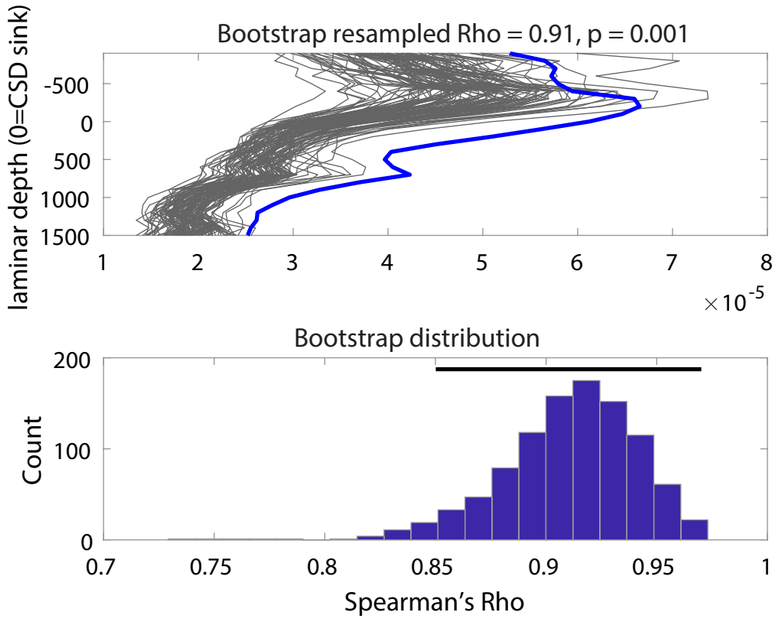


Supplemental Figure 7

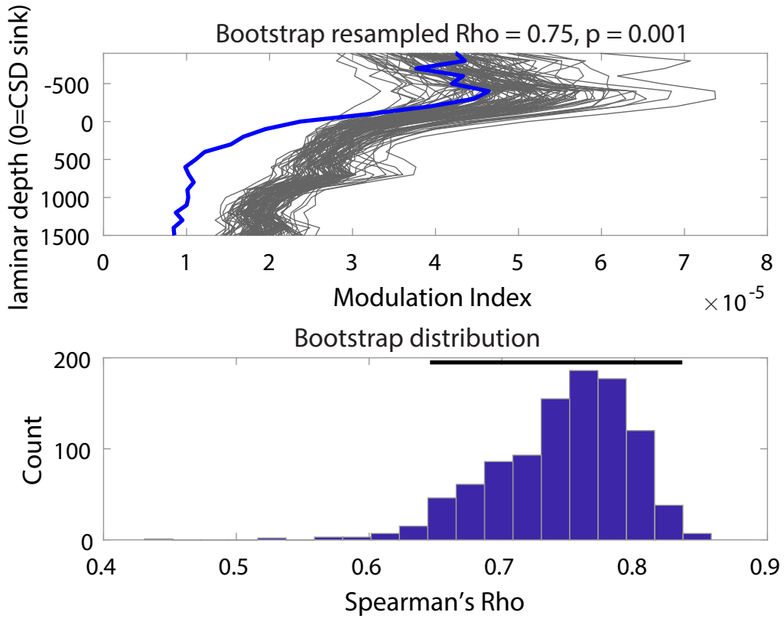


Supplemental Figure 8

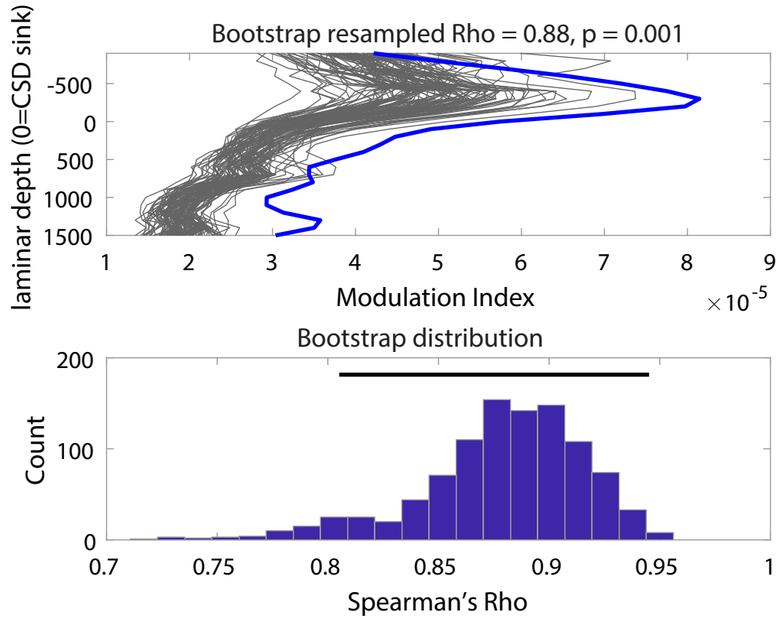
VLPFC (N=23)



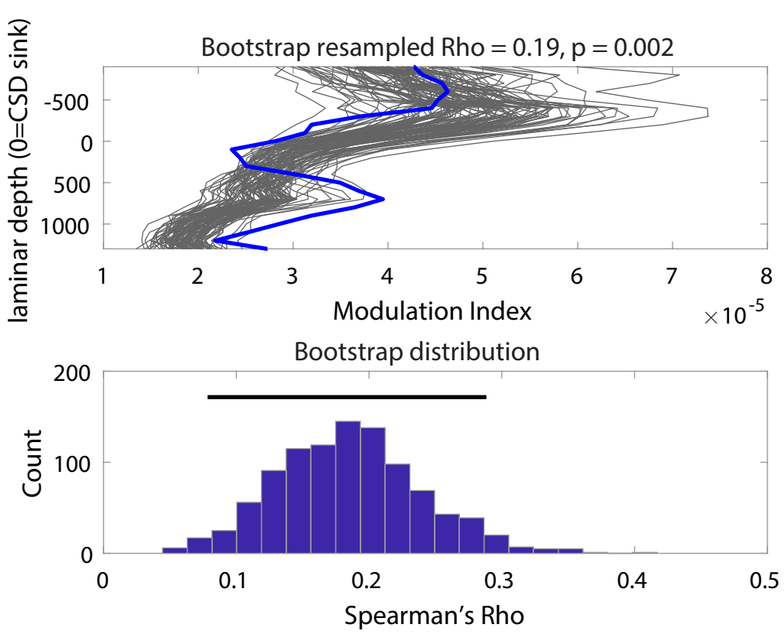
PMd (N=20)



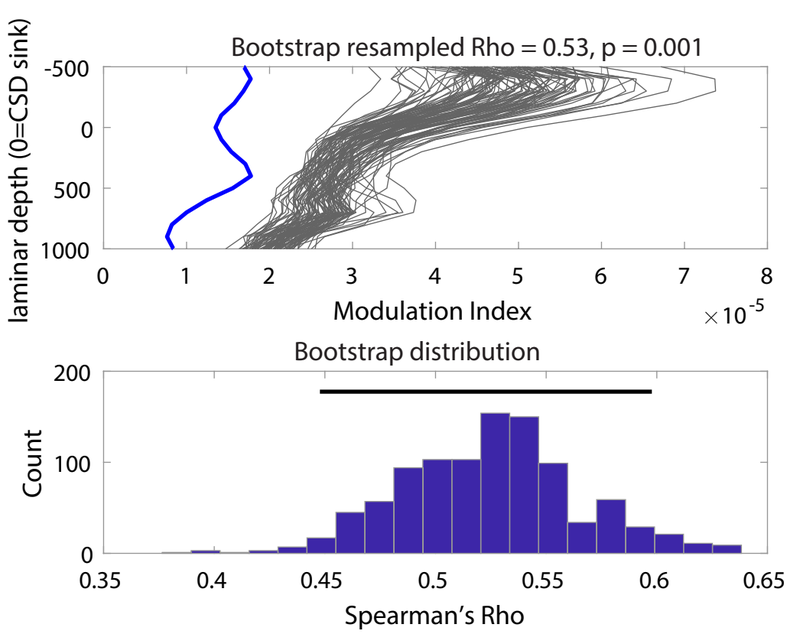
8A (N=8)



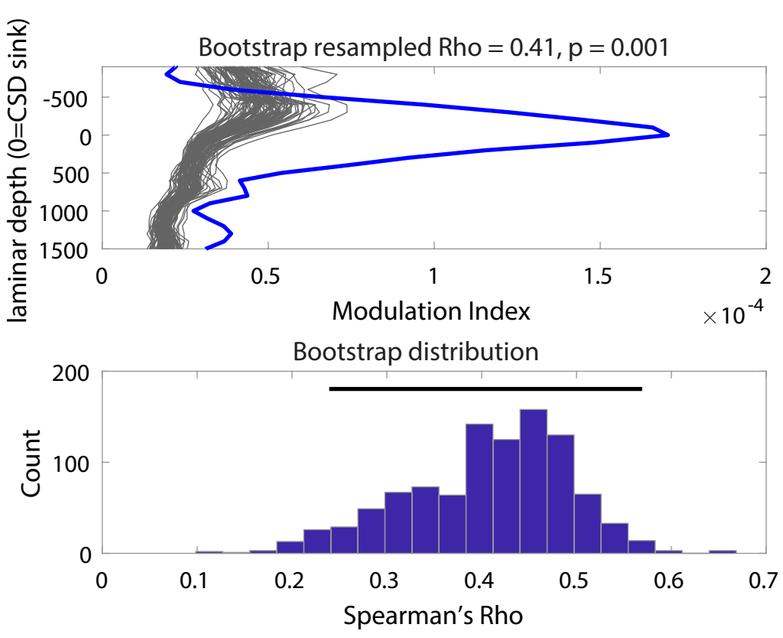
ACC/SMA (N=4)



8B (N=3)



DLPFC (N=2)

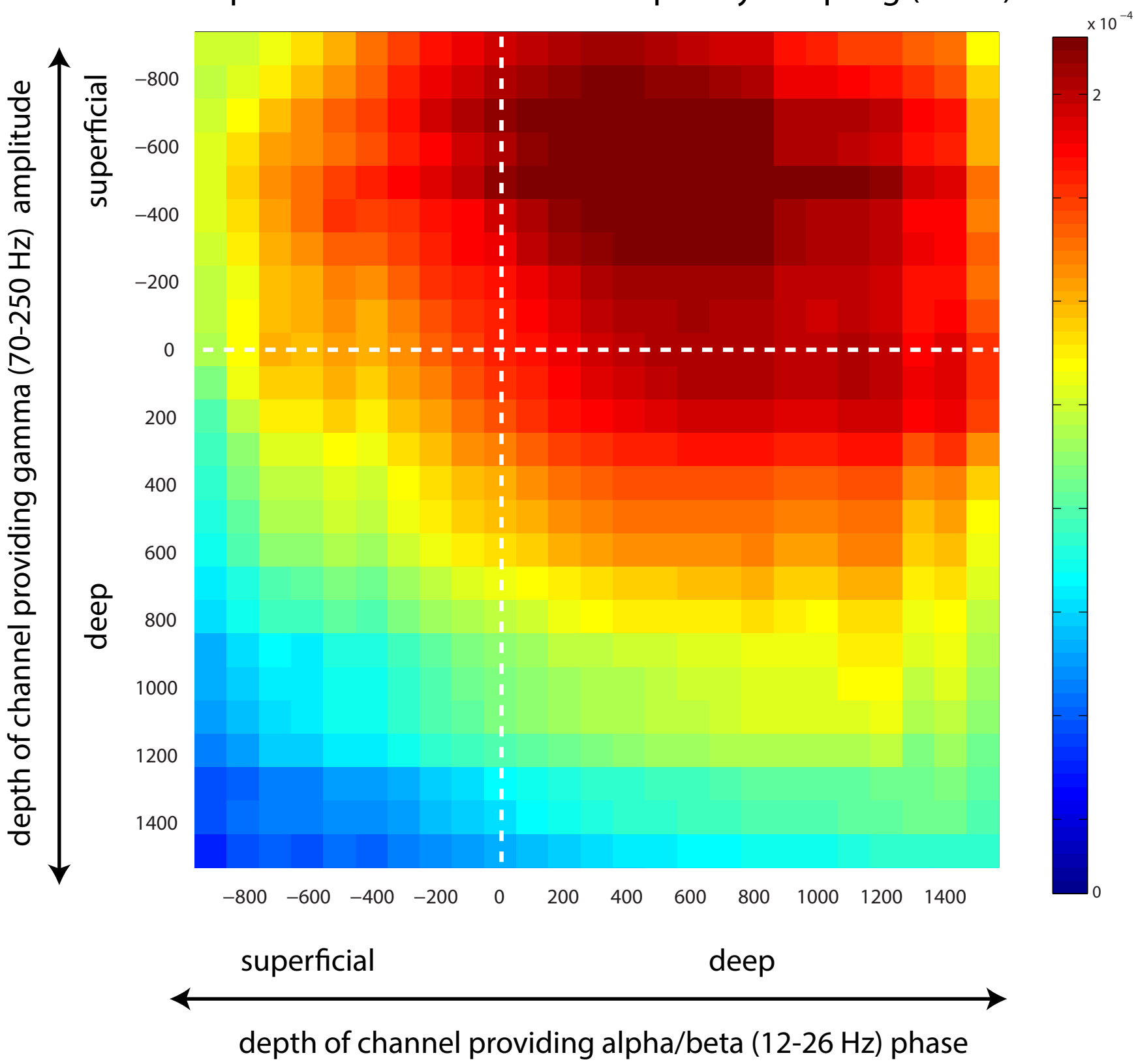


Legend

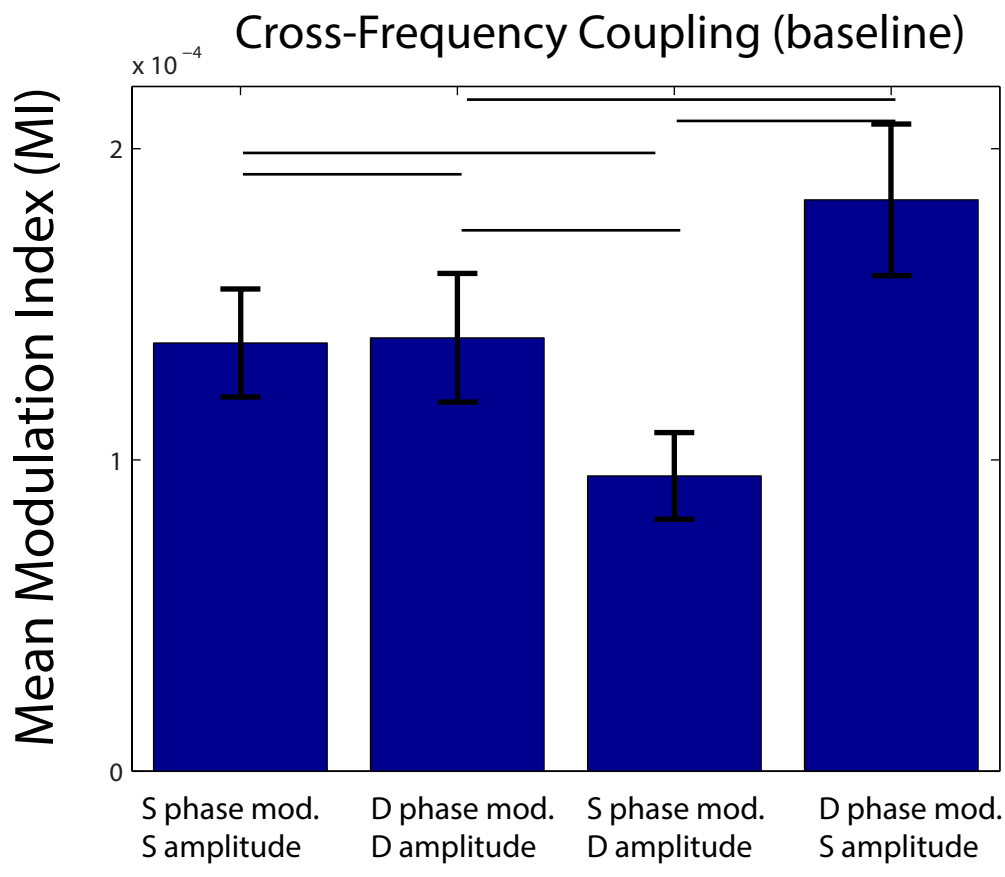
- Cross-Frequency Coupling (resampled average of all other areas)
- Cross-Frequency Coupling (average of each area)
- 95% Confidence Interval for Spearman's Rho

Supplemental Figure 9

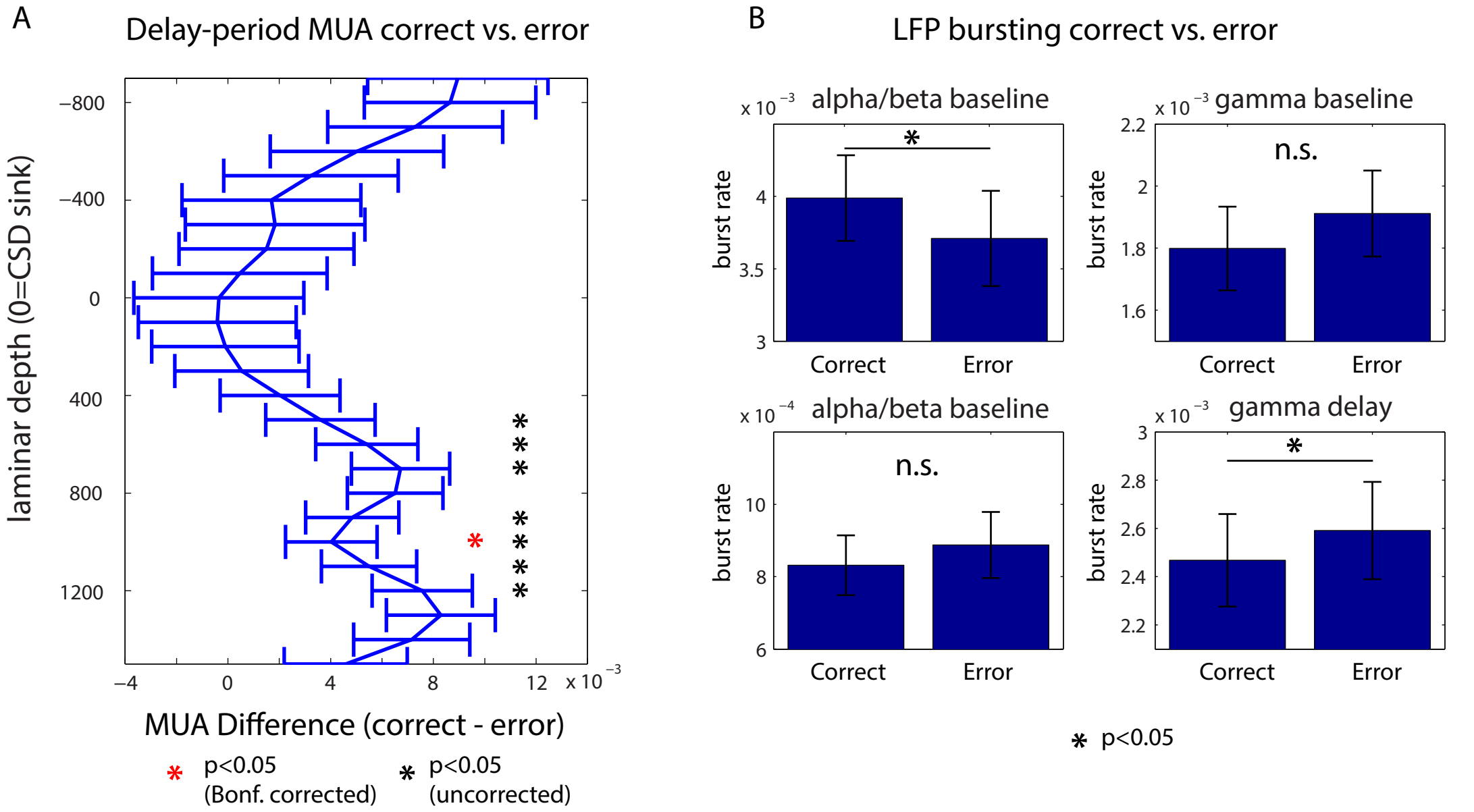
A



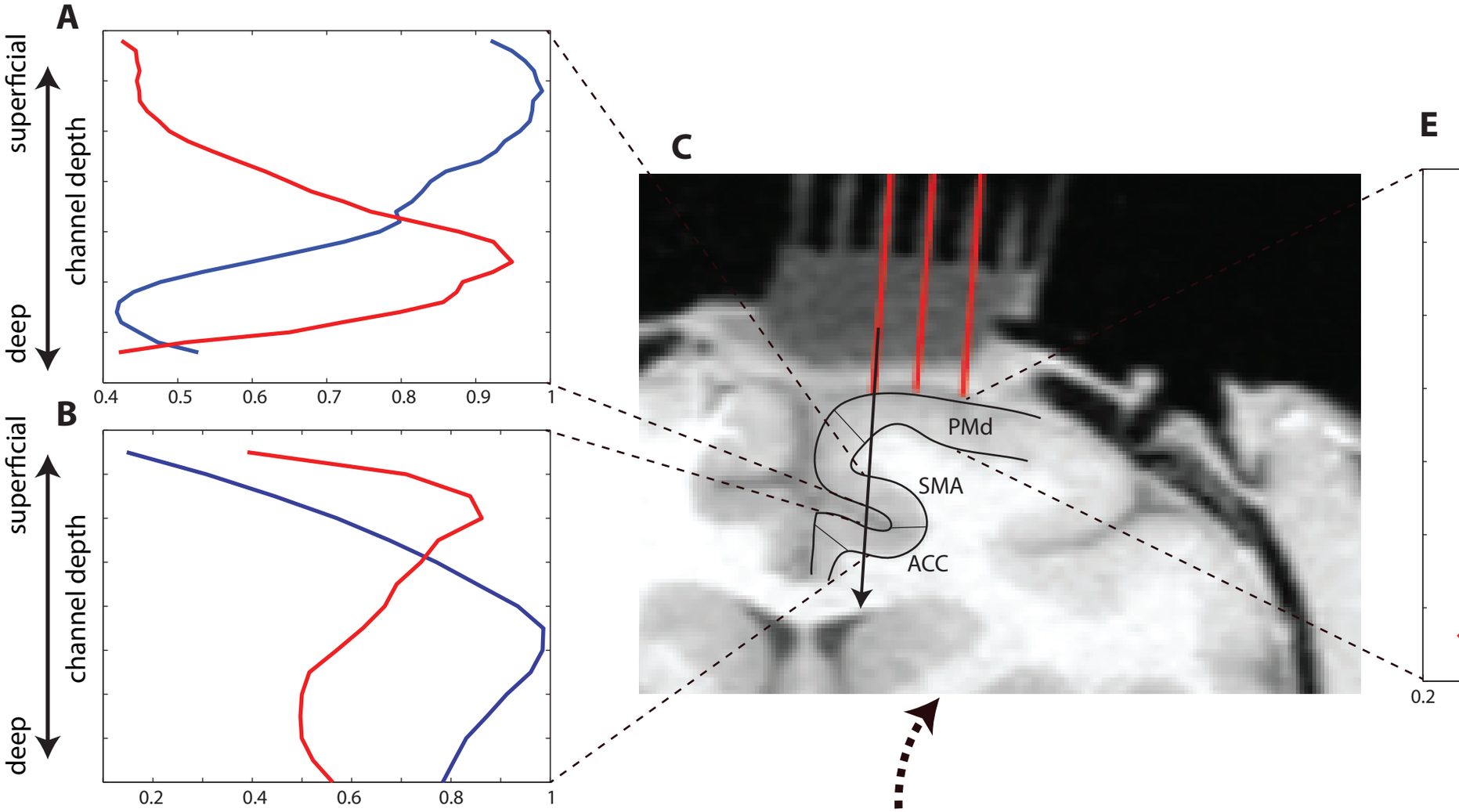
B



Supplemental Figure 10



Supplemental Figure 11



Supplemental Figure 12