Supporting Information Corrected September 14, 2015

Multiple Component Networks Support Working Memory in Prefrontal Cortex

David A. Markowitz, Clayton E. Curtis and Bijan Pesaran

Table of Contents

Supplemental Experimental Procedures

Experimental Preparation

Two adult male rhesus macaques (*Macaca mulatta*) participated in the study (Monkey A and Monkey S, 9.5 kg and 8.4 kg, respectively at the start of the experiments). Both animals had been used previously in other experiments studying eye movements, and identical training protocols were used for both animals (see below). Prior to behavioral training, each animal was instrumented with a head-restraint prosthesis to allow fixation of head position and tracking of eye position. Each monkey was behaviorally trained for several weeks in an unlit sound-attenuated electromagnetically shielded room (ETS Lindgren). Following behavioral training, we implanted a low-profile recording chamber (Gray Matter Research, MT) in a craniotomy made over the right pre-arcuate cortex of each animal using image-guided stereotaxic surgical techniques (Brainsight, Rogue Research, Canada). A semi-chronic microelectrode array microdrive (SC32-1, Gray Matter Research, MT) was then inserted into the recording chamber and sealed (see below). Prior to each behavioral session, we advanced electrodes in 15 μm increments until up to 30 neurons were isolated. Then we recorded from isolated neurons while each monkey performed up to 500 trials of randomly interleaved memory- and visually-guided delayed saccades to one of eight targets for a liquid reward (1). Eye position was constantly monitored with an infrared optical eye tracking system sampling at 120 Hz (ISCAN). After the completion of experiments and all 32 electrodes had been advanced into white matter, we registered previously measured absolute cortical depths to a common zero point across the array using an iterative algorithm (2).

Behavioral Tasks

Each monkey performed a **memory-guided oculomotor delayed response (mODR) task** to

one of eight isoeccentric targets for a liquid reward (**Fig. 1A**) (1). All trials began with the illumination of a central fixation target, on which the animal needed to maintain fixation for a baseline period (500 - 800 ms). After the baseline period, a spatial cue was flashed for 300 ms at a peripheral location to indicate the target of the saccade. After a delay period (typically 1000-1500 ms for both animals), the central fixation square was extinguished, providing the Go signal for the animal to saccade to the remembered location of the peripheral target. Within 100 - 150 ms of the saccade, the spatial cue reappeared and the animal had to maintain fixation on the cue for an additional 300 ms. A fluid reward was then delivered. On each trial, a spatial cue was presented at one location on a grid of eight locations spaced 10º around the central fixation target. Spatial cue locations were interleaved trial-by-trial in equal proportions.

A trial was aborted if the monkey failed to align its gaze within 2° of the center of the fixation target before the Go command or within 2° of the center of a spatial cue target following an initially correct saccade after Go. When an abort was detected, all visual stimuli were extinguished immediately, no reinforcers were delivered, and the trial was restarted after a 1200 - 1800 ms intertrial interval. Both monkeys rarely aborted trials (4% for Monkey A, 5% for Monkey S). Aborted trials were excluded from further analyses.

Each monkey also performed a **visually-guided oculomotor delayed response (mODR) task** (1). In this task, events proceeded as in the memory-guided task, except that the spatial cue was not flashed, remained illuminated throughout the delay period and saccade, and was extinguished after the trial was completed. Memory-guided task trials and visually-guided task trials were randomly interleaved trial-by-trial with equal proportions.

During each behavioral session, Monkey A performed 240-300 delayed saccades, and Monkey S performed 400-500 delayed saccades. Data reported here were collected after at least 3

weeks of training on the saccade tasks. Monkey A's reaction times (RTs) differed significantly across the mODR (211 ± 0.6 ms s.e.m.) and vODR tasks (171 ± 0.5 ms s.e.m.) ($p<0.05$, t-test), while Monkey S's reaction times did not differ across the mODR (189 ± 0.3 ms s.e.m.) and vODR tasks (188 ± 0.3 ms s.e.m.).

Microelectrode Microdrive Design

All data reported here were obtained using a semi-chronic microelectrode array microdrive, SC32-1 (Gray Matter Research, MT) (2–5)**.** The SC32-1 is a modular, replaceable micromanipulator system capable of independent bidirectional control of 32 microelectrodes. The system is designed to be semi-chronically implanted within a recording chamber system, and can be secured with acrylic for additional protection. Electrodes are spaced by 1.5 mm. The system was implanted for approximately two months in each animal and permitted long-term recordings of neuronal activity. The SC32-1 utilizes a screw-driven mechanism to bi-directionally control electrode position along a single axis with a range up to 20 mm. Each actuator consists of a lead screw, an eccentric brass shuttle mounted to the electrode, and a compression spring. Following implantation, single electrodes could be moved with an accuracy of approximately 15 μm and allowed sufficient control to stabilize isolated recordings of spiking activity from individual neurons.

Recording Protocol

In each animal, we advanced electrodes to maximize the yield of isolated single unit recordings during each recording session. The strategy employed was refined over the course of the experiments but followed the same general outline. At the time of implantation, the initial position of each recording electrode was recessed approximately 1 mm within the drive. Electrodes were advanced through a silastic membrane in the recording chamber, the dura

mater and pia before entering the cortex. After piercing the dura, each electrode was advanced in sequential 15 μm increments 10 minutes apart, to give the electrode time to settle in the tissue. After each movement iteration and settling interval, we obtained a 60 second recording while the animal sat quietly with the lights on. Action potentials were first recorded at a median depth of approximately 3 mm beyond their initial position (2.23 mm in Monkey A; 3.04 mm in Monkey S). Electrodes were gradually advanced across sessions (mean 34 μm/day in Monkey A; 100 μm/day in Monkey S) until action potentials were no longer present, indicating passage into white matter. We obtained neural recordings by advancing the electrodes up to a median distance of 6 mm from their initial position. Neural recordings during the performance of the behavioral task were first obtained in Monkey A when clearly isolated single unit activity was present on all channels. In Monkey S, neural recordings were also obtained at more superficial depths, before isolated single unit activity was present.

Data Acquisition

Eye position was constantly monitored with an infrared optical eye tracking system sampling at 120 Hz (ISCAN). Eye positions were digitized at 1 kHz. Visual stimuli were presented on an LCD screen (Dell Inc) placed 34 cm from the subjects' eyes. The visual stimuli were controlled via custom LabVIEW (National Instruments) software executed on a real-time embedded system (NI PXI-8184, National Instruments). Neural recordings were made with glass-coated tungsten electrodes (Alpha Omega, Israel) with impedance 0.7-1.5 MΩ measured at 1 kHz (Bak Electronics, MD). Neural signals were preamplified (10x gain; Multichannel Systems, Germany), amplified and digitized (16 bits at 30 kHz; NSpike, Harvard Instrumentation Lab), and continuously streamed to disk during the experiment (custom C and Matlab code).

Local field potential (LFP) waveforms were computed from the broad-band activity by median-filtering the raw, broad-band recording with a 1.5 ms window to suppress large amplitude

spiking events, low-pass filtering at 300 Hz and down-sampling at 1 kHz. Multiunit activity (MUA) waveforms were computed by high-pass filtering the raw data at 300 Hz and maintaining the original 30 kHz sampling rate. Single-unit activity (SUA) was isolated by thresholding MUA waveforms at 3.5 standard deviations below the mean, performing a principal component analysis of putative spike waveforms, over-clustering these waveforms in PCA space using k-means and then merging clusters based on visual inspection.

Data Analysis

We investigated distributed firing rate dynamics during WM using two dimensionality reduction approaches: principal component analysis (PCA), and a targeted dimensionality reduction (TDR) approach, which maps population activity into a low-dimensional state space that captures variability due to target location and delay type (6). PCA provided a compact and informative (yet fundamentally qualitative) description of our neural data that enabled us to form hypotheses about task-related differences between populations. Then in separate analyses of single unit responses, we tested these hypotheses through the use of permutation testing to quantify the significance of each unit's spatial tuning and selectivity for the memory or visual delay condition. We then used TDR to confirm that task-selective single unit responses are representative of task-related modes at the population level. We estimated the noise correlation for each pair of neurons under each delay condition by converting spike counts during the time interval of interest on each trial to target-specific z-scores, and then estimating Pearson's correlation coefficient between the z-score pairs across all trials (7). Finally, we estimated spike-field coherence (SFC) as a function of frequency and time using multitaper spectral estimation (8).

Assessing Recording Stability

After implanting the microdrive in each animal, we monitored the variance of LFPs and MUA across the array over time for evidence of the brain's inflammatory response. In Monkey A, we monitored LFP and MUA variance on all channels while holding electrodes at a constant depth for two weeks after implantation, but did not observe any systematic changes over time. While advancing electrodes in both animals, we continued to observe robust LFP and MUA signals across a majority of the 59 post-implantation days in Monkey A and 68 days in Monkey S.

Single Unit Isolation

Spike events were extracted and classified from the broadband activity using custom Matlab code (The Mathworks) during each recording session and resorted offline. To account for nonstationarity in the recordings, spike classification was done on a 100 s moving window, and clusters were tracked across windows. Occasionally there were periods when clusters were not isolated. Trials during those periods were marked, and these data were not subject to further analysis.

Depth Registration

After the completion of experiments and all 32 electrodes had been advanced into white matter, we registered previously measured absolute cortical depths to a common zero point across the array (i.e. the cortical surface) using an iterative algorithm, previously described in (2). First, for each electrode penetration, we generated a high spatial resolution map of LFP variance as a function of depth by estimating the LFP variance during the 60-second recording after each 15 μm movement iteration while the animals were sitting quietly with the lights on. Then we applied a variance stabilizing transformation to each channel's LFP variance depth profile to correct for minor differences in scale due to variable electrode impedances across the array. This

transformation involved calculating the logarithm of the LFP variance at each depth and then rescaling log variance data to range from 0-1 on each channel. Next, we identified the alignment for each channel pair that minimized the pairwise Euclidean distance between their normalized variance depth profiles. To find this optimal alignment, we calculated the pairwise Euclidean distance at relative depth offsets ranging from -1 mm to +1 mm. The offset with minimal pairwise Euclidean distance was labeled optimal. To determine the best depth offset for a given channel, we estimated the optimal offset with respect to each of the other 31 channels, and then calculated the mean of these values. Then we shifted the variance depth profile for each channel by half of its optimal mean offset. This shift operation was performed in batch mode, with all electrodes shifted simultaneously once all mean offsets had been calculated. This entire procedure was repeated iteratively until no further shifts were required. Results were inspected by calculating the array-averaged LFP variance depth profile and looking for an inflection point characteristic of a change in activity at a specific depth. We labeled this inflection point zero cortical depth in both animals, and confirmed its correspondence to the top of cortex by plotting the depths of all observed spiking units across the array. In both animals, the inflection point occurred within ±200 μm of the depth of the first observed spikes on all channels. Following depth registration, we observed a small number of electrode depth profiles in Monkey S with spiking activity that spanned >2.5 mm in depth, which exceeds the largest cortical thickness observed in areas 8 and 46 (9). Due to the 3D geometry of cortex, these electrodes likely entered cortex at angles slightly different from 90 degrees to the surface, leading to slightly "stretched" spike depth profiles. We corrected this problem by manually rescaling registered depths by 0.8 - 1.0 on affected electrodes to bring the span of spiking activity closer to 2.5 mm.

Sulcal Electrode Identification

We identified putative sulcal electrodes as those that required electrode movement by at least 3 mm below the dura before extracellular action potentials were recorded. This step identified ~10

sulcal candidate electrodes in each animal. We confirmed or rejected these candidates by tracing a three-dimensional region-of-interest (ROI) in each animal's MRI, using the chamber registration to map electrode trajectories through the ROI, and identifying electrodes that projected down sulcal banks. To compensate for the +/- 1 mm positional error and +/- 20 degree rotational error of the chamber registration, we repeated this procedure using all possible registrations within this range of offsets, in 0.1 mm and 1 degree increments. We rejected candidate registrations for which the predicted depth of first spike contact disagreed with the observed depth of spike contact by more than 2 mm on any electrode. Finally, we identified the candidate registration with the greatest overlap between the ROI and observed spike locations. This optimal registration for each animal was then used to confirm or reject putative sulcal electrodes. In Monkey A, 4 channels were rejected as sulcal to give a total of 28 electrodes for further analysis. In Monkey S, 2 channels were rejected as sulcal and an additional two channels were rejected due to lack of any observed neural signals to give a total of 28 electrodes for further analysis. All results presented in this study hold regardless of whether putative sulcal electrodes were included or excluded from analysis.

Dimensionality Reduction Analyses

To study the low-dimensional dynamics of population activity, first we used principal component analysis (PCA) to identify eigenmodes of population activity that compactly describe the responses of all 746 isolated neurons across all eight targets during both tasks. We generated a time-dependent firing rate vector for each neuron during each trial by counting spikes in sequential 50 ms bins and smoothing data with a 20 ms Gaussian kernel. Each firing rate trace was a concatenation of three data alignments: -300 to +1,000 ms around Cue onset, -500 to 0 ms around the Go command, and -200 to +500 around Saccade onset. Next, we calculated each neuron's trial-averaged response to repetitions of each target during each task and then

normalized with respect to the mean and variance of the firing rate during a 300 ms baseline interval before Cue onset. Then, for each neuron and target, we appended visual task data to memory task data, resulting in 746 x 8 distinct firing rate traces. Finally, we diagonalized the covariance matrix of these data to discover the eigenvalues (principal components) and associated eigenmodes that best explain both the common and independent sources of variance in the data (10). In the context of multivariate data, this analysis produces a new coordinate system in which the first coordinate accounts for as much variance in the data as possible, the second coordinate for as much of the remaining variance as possible, etc.

To exclude independent sources of variability in the data and discover only the common-mode components of population activity that are driven by task variables, we used a technique that combines linear regression with "targeted dimensionality reduction" (TDR) to investigate population dynamics during WM (6). We briefly describe our application of this technique here, and provide the complete mathematical details of this method in a subsequent section titled "Mathematical Appendix: Targeted Dimensionality Reduction." To prepare data for this analysis, first, we generated a time-dependent firing rate vector for each neuron during each trial by counting spikes in sequential 50 ms bins and smoothing data with a 40 ms Gaussian kernel. All firing rates were then normalized with respect to the mean and variance of the firing rates across all units, times and trials. After firing rate estimation, we enforced two constraints to guarantee that all trials in our analysis had the same duration: first, where necessary, delay activity was compressed in time (by rescaling data in the time domain) to span 1 s, and activity between the time of the Go command and saccade onset was compressed or dilated in time to span 200 ms. Then we regressed each neuron's firing rate over the two task variables that were manipulated in our experiments: target location and memory/visual delay context. Although eight targets were presented during experiments, for simplicity, we only plot responses to two specific targets in this analysis: the leftmost contralateral target, and the rightmost ipsilateral

target. Similar results are achieved by plotting responses to nearest neighbor targets of the most contralateral and ipsilateral locations. Therefore, our findings are not specific to one pair of locations only.

As detailed in the **Mathematical Appendix**, we then used PCA to discover a low-dimensional subspace that explains most of the variance in activity by all 746 isolated neurons across all times and task conditions. We focused our subsequent analysis on the subspace spanned by the first N=8 principal components (**Fig. S4A**). Next, we assembled the previously calculated regression coefficients for all neurons into a population matrix and "de-noised" the coefficients by projecting into the PCA subspace. (This procedure was applied to population activity at each time point, leading to a different set of de-noised coefficients at each time point. The set with largest L_2 norm was selected as the "best" set of coefficients for each variable.) Finally, we orthogonalized the “best” de-noised regression vectors. Population activity was then projected onto each of the two vectors during each task condition to map activity into a state space that maximally captures variability due to the associated task variable (**Fig. S4B,C**). We used cross-validation to confirm that neural trajectories in state space do not change when regression-vectors are estimated from and projected onto random, non-overlapping 50% subsets of trial data (**Fig. S4D**).

Spike Waveform Classification

Previous work has shown that certain fast-spiking (FS) inhibitory neurons in primate PFC are distinguishable from regular-spiking (RS) excitatory cells by their baseline firing rate, spike waveform shape, and task-responsiveness (11, 12). Due to these functional differences, we classified the units in our spike database as FS or RS using a conjunction of criteria. First, for each unit, we calculated the mean spike waveform across all samples, and then estimated two

features: the peak-to-peak time (P2P), defined as the time in milliseconds between the two maximal values on either side of the action potential trough; and the trough-to-peak time (T2P), defined as the time in milliseconds between the action potential trough and subsequent voltage peak. We also estimated the firing rate of each unit during the 500 ms Baseline interval before Cue onset, averaged across all ODR trials. All units were initially classified as RS. Then, we classified each unit as FS if P2P < 0.54 ms and the baseline firing rate exceeded 8.7 spikes/second, the sample mean across all recorded units (11), or if T2P < 0.25 ms (13–15). Applying this set of quantitative criteria labeled 99 units as FS (13.3% of total) and 647 units as RS (86.7% of total) in our combined database across both monkeys. The population of FS units is qualitatively distinguishable from RS units as a mostly distinct band in T2P vs P2P feature space, which corresponds to clear differences in the mean spike waveforms across these populations (**Fig. S3**).

Tuning Z-Score Calculation

We quantified the significance of each unit's spatial tuning using a tuning z-score (16). First, we estimated the mean firing rate, f_i, of each unit across all presentations of each Cue location, (indexed by *i*). Next, we assigned an angular displacement, φ_i, to each Cue location and then calculated the first trigonometric moment, R_m, of each unit's response across all eight angles, as follows:

$$C = \sum (f_i \cdot \cos \varphi_i) \quad (1)$$

$$S = \sum (f_i \cdot \sin \varphi_i) \quad (2)$$

$$R_m = \sqrt{[\,(C/\sum f_i)^2\,] + [\,(S/\sum f_i)^2\,]} \quad (3)$$

To establish a null distribution, we shuffled target labels across trials and repeated this moment

estimation procedure 10,000 times. The p-value, or fraction of shuffled moment values that exceeded the true moment of the data, was then passed to a normal inverse cumulative distribution function (*norminv* in Matlab) to obtain the tuning z-score. Therefore, the tuning z-score is calculated with respect to shuffled data at the same moment in time as the test data. Firing rates were estimated from 300 ms windows during the delay epoch of each task, immediately before the Go command. A unit was labeled as tuned if its tuning z-score exceeded 1.65 (one-tailed t-test) for trials of either the vODR or mODR task. A unit was labeled with inverted tuning if its preferred stimulus response during the delay was less than the firing rate during the baseline (17). Across both monkeys, we identified 365 RS units with positive tuning, 92 RS units with inverted tuning, 59 FS units with positive tuning, and 19 FS units with inverted tuning. This method was adapted to calculate LFP spatial tuning at each frequency during the last 300 ms of the memory delay (**Fig. S6A**) by substituting the trial-averaged power spectrum for firing rates.

Task Selectivity Analysis

Among the 365 RS neurons with positive spatial tuning during the delay, we quantified the task selectivity of each unit's preferred target response using a permutation test**.** First, we estimated the trial-averaged firing rate of each neuron in response to its preferred target during the last 300 ms of the delay interval of both tasks. Then, we compared the true difference in firing rates across tasks (mODR rate – vODR rate) to a resampled difference estimate obtained by merging trials from both tasks, randomly resampling mODR and vODR trials from the merged distribution 10,000 times, and recalculating the difference in mean rates for the resampled trials during each iteration. The p-value, or fraction of resampled rate differences that exceeded the true rate difference, was then passed to a normal inverse cumulative distribution function (*norminv* in Matlab) to obtain the task selectivity z-score. Units with a selectivity z-score < -1.65 (1-tailed t-

test) were classified as “early storage” neurons, and units with a z-score > +1.65 (1-tailed t-test) were classified as “late storage” neurons. All non-selective units with z-scores between (-1.65, +1.65) were labeled as putative response neurons. This method was adapted to calculate LFP task selectivity at each frequency during the last 300 ms of the delay (**Fig. S6B**) by substituting the trial-averaged power spectrum for firing rates.

Reaction Time Analysis

We grouped mODR trials from each monkey into two categories containing the fastest 50% and slowest 50% of RTs after the Go command, respectively. Then, for each unit in the early storage, late storage and response populations, we selected preferred target trials from each RT category and estimated firing rates during the last 500 ms of the delay. In 49% of cases, fewer than 10 trials were available for firing rate estimation in at least one RT condition. Subsequent decimation analysis revealed a strong sample dependence of trial-averaged firing rate estimates in most of these cases, indicating that splitting trials into multiple RT categories tends to result in undersampled mean firing rate estimates for single neurons. Therefore, we did not perform a two-way ANOVA of unit class (early storage, late storage, response) and RT (fast or slow) on firing rate for individual neurons. Instead, we pooled all trials from each RT category across each unit class and then used permutation testing with 10,000 iterations to identify significant differences in firing rate across RT categories on a population basis. This method treats every trial’s firing rate estimate as a representative sample from a single population, rather than from a single neuron. In this manner, any observed difference in firing rate across RT conditions can be interpreted to reflect a property of the population.

Error Trial Analysis

We distinguish error trials (when the monkey received the Go command and then saccaded to a location outside the ±2^{o} target bounding box) from aborted trials (when the monkey saccaded prior to receiving the Go command, or received the Go command but failed to saccade, or saccaded to the correct location after Go but broke fixation prematurely). The fraction of error trials was 2.6% during the mODR task (373 error, 13,894 correct), and memory saccade error endpoints were separated from the instructed target by a median angle of 15.4^{o}, suggesting a critical failure of WM before these rare mistakes. Individually, Monkey A showed a 1.4% error rate and Monkey S showed a 4.0% error rate during the mODR task. Monkey A's RTs during error trials (193 ± 24.8 ms s.e.m.) were significantly faster than RTs during correct trials (211 ± 0.6 ms s.e.m) ($p<0.01$, permutation test). By contrast, Monkey S's RTs during error trials (279 ± 7.8 ms s.e.m.) were significantly slower than RTs during correct trials (189 ± 0.3 ms s.e.m.) ($p<0.001$, permutation test).

Spike-Field Coherence Analysis

We estimated spike-field coherence (SFC) as a function of frequency and time using multitaper spectral estimation (8, 18) with 4 Hz smoothing (for frequencies from 4-13 Hz) or 10 Hz smoothing (for frequencies from 14 to 100 Hz), and an estimation window spanning the last 1000 ms of the delay period. To study SFC between the late storage network and fields on other electrodes, first we selected all pairs that included a late storage unit on one electrode and any class of delay-tuned unit with the same preferred target on a different electrode, which was labeled the field electrode. This selection criterion was designed to increase the likelihood that field potentials would reflect processing of the late storage unit's preferred target at other PFC recording sites. In contrast to our single unit data, we did not assign class labels to LFPs because these signals tended to exhibit memory delay-selectivity in PFC regardless of which

class of unit was recorded nearby (**Fig. S6B**). Next, we identified trials when the preferred target of the late storage unit was presented, and then estimated the coherence magnitude after pooling all late storage unit spike data and all field data from the field electrodes across these trials. To estimate the standard error of our estimator, we repeated the coherence estimate 10,000 times after randomly resampling data with replacement. To establish a null distribution, we shuffled trial labels within each unit pair and repeated the coherence estimation procedure 10,000 times. Raw coherence values were converted to z-scores by subtracting the mean and then dividing by the standard deviation of the null distribution. This procedure was repeated for the early storage, late storage and response populations. We identified significant differences in coherence across reaction time conditions using a permutation test with 10,000 iterations, and rejected spuriously significant frequency bands (cluster correction) by comparing the length of each contiguous range of significantly coherent frequencies with a null distribution that was obtained from shuffled data (19). If the length of a significant band in unshuffled data failed to exceed 95% of the significant band lengths in shuffled data, it was deemed not significant.

Noise Correlation Analysis

"Noise correlation," or spike count correlation, between two neurons describes any correlated variability in firing rates across repeated presentations of the same stimulus that remains after subtracting out each neuron's mean response to the stimulus, i.e. after controlling for signal correlation. We quantified noise correlations between unit pairs that were drawn from the same response class (early storage, late storage or response) and satisfied a relative tuning constraint (such as a 45 degree maximum difference in preferred target angle). To estimate noise correlation for each pair of neurons during the mODR task, we counted spikes by each neuron during the last 300 ms of the memory delay on each trial, converted these responses to target-specific z-scores by subtracting the mean response across repeated target presentations

and dividing by the standard deviation, and finally estimated Pearson's correlation coefficient between the firing rate vectors across all trials, spanning all targets (7). This process was repeated for vODR trials to estimate noise correlation during the visual delay. We confirmed that the correlation estimates reported here were independent of firing rate by quantifying the relationship between geometric mean firing rate and the variance of noise correlation estimates (R^2<0.05 for all mODR and vODR task comparisons).

Relationship of Labeled Units to Classical Visual, Visuomovement and Movement Units

For the interested reader, ***Table 1*** below summarizes how the three classes of persistently active units identified in our study relate to the classical heuristic labeling of PFC neurons as "visual," "visuomovement" or "movement" (20). We assigned classical labels by applying the criteria specified in (21): To classify neurons as visual, visuomovement and movement, we measured spike counts within specified windows. Visual responses were measured between 50 and 150 ms after Cue onset. Baseline activity was measured between 150 ms and 0 ms before Cue onset. Movement responses were measured between 100 ms before and 20 ms after the initiation of the saccade. Premovement activity was measured between 350 ms and 200 ms before the initiation of the saccade. A neuron was classified as visual if the visual response was significantly greater than baseline activity ($p < 0.05$, permutation test) in at least one target location and the movement response was not significantly greater than the premovement activity at any target location. Accordingly, a neuron was classified as movement related if the movement response was significantly greater than the premovement activity ($p < 0.05$) for saccades to at least one target location. Visuomovement neurons displayed significant visual and movement responses.

Table 1. *Summary of our unit labels in relation to classical PFC labels.*

Unit totals are presented by waveform class, i.e. N_{RS} (N_{FS})

	Early Storage	*Late Storage*	*Response*	*Other*	*Total*
Visual	23 (3)	22 (3)	56 (10)	73 (10)	174 (26)
Visuomovement	28 (8)	39 (7)	69 (14)	67 (21)	203 (50)
Movement	5 (1)	18 (4)	35 (1)	53 (2)	111 (8)
Other	3 (1)	14 (2)	40 (3)	102 (9)	159 (15)
Total	59 (13)	93 (16)	200 (28)	295 (42)	647 (99)

Mathematical Appendix: Targeted Dimensionality Reduction

To investigate the distributed dynamics of WM, we projected the population activity of all 746 isolated neurons onto a two-dimensional state space that maximally captures variability due to two task variables: target location (Target space) and memory/visual delay type (Context space). This analysis was supported by a technique that combines linear regression with targeted dimensionality reduction (6), the details of which are reproduced below. Modifications to this procedure that are specific to our task design are indicated, where appropriate.

1. Linear Regression

We used multi-variable, linear regression to determine how the two task variables (target location and memory/visual delay context) affect the responses of each recorded unit. We first z-scored the responses of a given unit by subtracting the mean response from the firing rate at each time and in each trial, and by dividing the result by the standard deviation of the responses. Both the mean and the standard deviation were computed by combining the unit's responses across all trials and times. We then describe the z-scored responses of unit *i* at time *t* as a linear combination of two task variables:

$$r_{i,k}(t) = \beta_{i,t}^{(1)} \cdot Target(k) + \beta_{i,t}^{(2)} \cdot Context(k) + \beta_{i,t}^{(3)}. \quad (1)$$

where $r_{i,k}(t)$ is the z-scored response of unit *i* at time *t* and on trial *k*, Target(*k*) is the presented target on trial *k* (+1 for a contralateral target, -1 for an ipsilateral target), and Context(*k*) is the delay type on trial *k* (+1 for memory, -1 for visual). Although eight targets were presented during experiments, for simplicity, we only included trials of two specific targets in this analysis (the

leftmost contralateral target, and the rightmost ipsilateral target). Similar results are achieved by plotting responses to nearest neighbor targets of the most contralateral and ipsilateral locations. Therefore, our findings are not specific to one pair of locations only.

The regression coefficients $\boldsymbol{\beta}_{i,t}^{(v)}$ for v = 1 to 3, describe how much the trial-by-trial firing rate of unit i, at a given time t during the trial, depends on the corresponding task variable v. Here, and below, v indexes the two task variables, i.e. Target (v = 1) and Context (v = 2). The last regression coefficient (v = 3) captures variance that is independent of the two task variables, and instead results from differences in the responses across time.

To estimate the regression coefficients $\beta_{i,t}^{(v)}$ we first define, for each unit i, a matrix $\boldsymbol{F_i}$ of size N_{coef} x N_{trial}, where N_{coef} is the number of regression coefficients to be estimated (i.e. 3), and N_{trial} is the number of trials recorded for unit i. The first two rows of $\boldsymbol{F_i}$ each contain the trial-by-trial values of one of the four task variables. The last row consists only of ones, and is needed to estimate $\beta_{i,t}^{(3)}$. The regression coefficients can then be estimated as:

$$\boldsymbol{\beta}_{i,t} = \left(\boldsymbol{F}_i \boldsymbol{F}_i^{\boldsymbol{T}}\right)^{-\mathbf{1}} \boldsymbol{F}_i \, \boldsymbol{r}_{i,t} \qquad (2)$$

where $\boldsymbol{\beta}_{i,t}$ is a vector of length N_{coef} with elements $\beta_{i,t}^{(v)}$, v=1-3. Here and below we denote vectors and matrices with bold letters, and use the same letter (not bold) to refer to the corresponding entries of the vector or matrix, which in this case are indexed by v.

2. Population average responses

We constructed population responses by combining the condition-averaged responses of units that were mostly recorded separately, rather than simultaneously. We defined conditions based on the target location (contralateral or ipsilateral) and delay context (memory or visual). For each unit, trials were first sorted by condition, and then averaged within conditions. We then smoothed the responses in time with a Guassian kernel (σ = 40 ms). Finally, we z-scored the average, smoothed responses of a given unit by subtracting the mean response across times an conditions, and by dividing the result by the corresponding standard deviation. We define the population response for a given condition *c* and time *t* as a vector $x_{c,t}$ of length N_{unit} built by pooling the responses across all units for that condition and time. Therefore, the dimension of the state space corresponds to the number of units in the population.

3. Targeted dimensionality reduction

To understand the dynamics of PFC activity during WM, it is critical to identify the components of the population responses that are most tightly linked to the monkeys' behavior. Our ultimate goal is to define a small set of axes, within the state space of dimension N_{unit} defined by the activity of each unit, which independently account for response variance due to key task variables. The projection of the population responses onto these axes yields de-mixed estimates of the task-variables, which are mixed at the level of single neurons.

To define the axes of the subspace, we applied the "Targeted dimensionality reduction" approach developed by (6), consisting of three steps described in detail below. We start by using principal component analysis (PCA) to de-noise the population responses and focus our analyses on the subspace spanned by the first N_{pca} = 8 principal components (PCs). We then identify directions in this reduced subspace (the de-noised regression vectors defined below)

that together account for response variance due to 2 task variables (target and context). Finally, we orthogonalize the two identified directions to define axes that account for separate components of the variance due to the task variables.

4. Principal component analysis

We used PCA to identify the dimensions in state space that captured the most variance in the condition-averaged population responses. We first build a data matrix X of size N_{unit} x ($N_{condition}$ · T), whose columns correspond to the smoothed, z-scored population response vectors xc,t defined above for a given condition *c* and time *t* (section 1). $N_{condition}$ corresponds to the total number of conditions, and T to the number of time samples. The PCs of this data matrix are vectors v_a of length N_{unit}, indexed by a from the PC explaining the most variance to the one explaining the least. We use the first N_{pca} PCs to define a de-noising matrix $\boldsymbol{D}$ of size N_{unit} x N_{unit}:

$$\boldsymbol{D} = \sum_{a=1}^{N_{pca}} \boldsymbol{v}_a \boldsymbol{v}_a^T \ . \quad (3)$$

The de-noised population response for a given condition and time is defined by:

$$\boldsymbol{X}^{pca} = \boldsymbol{D}\,\boldsymbol{X} \ , \quad (3)$$

with $\boldsymbol{X}^{pca}$ also of dimension of size N_{unit} x ($N_{condition}$ · T). The overall contribution of the a^{th} PC to the population response at each time point *t* can be quantified by first projecting the population response onto that PC, and then computing the variance across all conditions of the projection, var($\boldsymbol{v}_a^T \boldsymbol{X}$) (**Fig. S4A**).

5. Regression subspace

We use the regression coefficients described in Equation 1 above to identify dimensions in state space containing task related variance. For each task variable v = 1-2 we first build a set of coefficient vectors $\boldsymbol{\beta}_{v,t}$ whose entries $\boldsymbol{\beta}_{v,t}(i)$ correspond to the regression coefficient for task variable v, time t, and unit i. The vectors $\boldsymbol{\beta}_{v,t}$ (of length N_{unit}) are obtained by simply rearranging the entries of the vectors $\boldsymbol{\beta}_{i,t}$ (of length N_{coef}) computed above (section 1). This re-arrangement corresponds to the fundamental conceptual step of viewing the regression coefficients not as properties of individual units, but as the directions in state space along which the underlying task variables are represented at the level of the population. Each vector, $\boldsymbol{\beta}_{v,t}$, thus corresponds to a direction in state space that accounts for variance in the population response at time t, due to variation in task variable v.

We de- noise each vector by projecting it into the subspace spanned by the first N_{pca} = 8 principal components:

$$\boldsymbol{\beta}_{v,t}^{pca} = \boldsymbol{D}\,\boldsymbol{\beta}_{\boldsymbol{v,t}}\,, \tag{4}$$

with the set of vectors $\boldsymbol{\beta}_{v,t}^{pca}$ also of length N_{unit}. We refer to these vectors as the 'de-noised' regression coefficients. This de-noising corresponds to removing from each vector $\boldsymbol{\beta}_{\boldsymbol{v,t}}$ the component lying outside the subspace spanned by the first N_{pca} = 8 PCs.

For each task variable *v*, we then determine the time, t_v^{max}, for which the corresponding set of vectors $\boldsymbol{\beta}_{v,t}^{pca}$ has maximum norm, and then define time-independent, de-noised 'regression

vectors':

$$\boldsymbol{\beta}_v^{max} = \boldsymbol{\beta}_{v,t_v^{max}}^{\boldsymbol{pca}} \text{ with} \quad (4)$$

$$t_v^{max} = \text{argmax}_t \left\| \boldsymbol{\beta}_{v,t}^{\boldsymbol{pca}} \right\| , \quad (5)$$

where each $\boldsymbol{\beta}_v^{max}$ is of dimension N_{unit}. Finally, we obtain the orthogonal axes of Target and Context (e.g. **Fig. S4B**) by orthogonalizing the regression vectors $\boldsymbol{\beta}_v^{max}$with the QR-decomposition:

$$\boldsymbol{B}^{max} = \boldsymbol{Q}\,\boldsymbol{R} \ , \quad (6)$$

where $\boldsymbol{B}^{max}$ = [$\boldsymbol{\beta}_1^{max}$ $\boldsymbol{\beta}_2^{max}$] is a matrix whose columns correspond to the regression vectors, ***Q*** is an orthogonal matrix, and ***R*** is an upper triangular matrix. The first two columns of Q correspond to the orthogonalized regression vectors Bv_orth, which we refer to as the 'task-related axes' of target and context. These axes span the same 'regression subspace' as the original regression vectors, but crucially each explains distinct portions of the variance in the responses.

To study the representation of the task-related variables in PFC, we projected the average population responses onto these orthogonal axes (**Fig. S4**):

$$p_{v,c} = \boldsymbol{\beta}_v^{\perp \boldsymbol{T}} \boldsymbol{X_c} \ , \quad (7)$$

where $p_{v,c}$is the set of time-series vectors over all task variables and conditions, each with length *T*. Furthermore, we have reorganized the data matrix, *X*, so that separate conditions are

in separate matrices, resulting in a set, X_c, of $N_{condition}$ matrices of size N_{unit} x T.

Supplemental Figure Legends

Figure S1. Tuning Curve Examples. Dots in each panel show the mean late delay response by a single unit following repeated presentation of 8 different targets during the mODR (red) and vODR (blue) tasks. Solid lines show the best fitting von Mises distribution for each task. (**A**) Early storage unit responses during the late delay interval (spanning the last 300 ms before the Go command). (**B**) Late storage unit responses during the late delay interval. (**C**) Response unit responses during the late delay interval.

Figure S2. Principal Component Analysis. (**A**) Eigenmodes corresponding to the largest five eigenvectors of the population firing rate data. Traces are color-coded by the mODR (red) and vODR (blue) components of each mode. Modes 3 and 5 exhibit pronounced task-selectivity. (**B**) Scatterplot of all neural responses that were used as source data for PCA after projection onto the 3rd and 5th eigenmodes. The lack of obvious clustering indicates that memory- and visually-selective responses do not cluster in PC-space.

Figure S3. Classification of FS and RS Units. (**A**) Scatterplot of Baseline Firing Rate and Peak-to-Peak Time (P2P) for n=746 units recorded across two monkeys. Vertical and horizontal lines define two thresholds (P2P < 0.54 ms and Baseline Rate > 8.7 sp/sec) that were used to classify units as FS (red) or RS (black). Shaded region defines units that were assigned to the FS category by these two criteria. (**B**) Scatterplot of Baseline Firing Rate and Trough-to-Peak Time (T2P) for the same units from (**A**). Vertical line defines an additional threshold (T2P < 0.25 ms) that was used to classify units as FS or RS. Shaded region defines units that were assigned to the FS category by this criterion. (**C**) Scatterplot of T2P and P2P spike features for the same units from (A), color-coded by FS or RS cluster assignment. (**D**) Mean spike waveforms across

all RS (black, n = 647) and FS (red, n = 99) units identified by this analysis.

Figure S4. Distinct Early and Late Storage Modes are Revealed By Targeted Dimensionality Reduction. (**A**) Percent of total variance explained by the first 20 eigenvectors calculated from the activity of all 746 units recorded across two monkeys. We de-noised population data by projecting all samples into the space of the 8 eigenvectors with largest eigenvalues. (**B**) Trajectories of all 746 isolated neurons in a two-dimensional state space that captures variability due to memory/visual delay type (Context space) and target location (Target space). All panels show contralateral target responses during mODR (red) and vODR (blue) trials. (**i**) Delay activity in the Target space reveals an early storage mode. (**ii**) Delay activity in the Context space reveals a late storage mode. (**C**) Neural trajectories in the Target and Context spaces following presentation of contralateral (solid) and ipsilateral (dotted) targets during mODR (red) and vODR (blue) trials. (**D**) Cross-validation of de-noised regression vectors estimated from two non-overlapping 50% subsets of trial data. These regression vectors map population activity onto nearly identical neural trajectories in state space.

Figure S5. Coding of Saccade Accuracy. **(A)** (Left) Angular error (Φ) and eccentricity error (ε) were quantified during mODR saccade error trials by measuring the indicated displacements of the saccade endpoint from the instructed Cue. (Right) Scatterplot of displacements during 373 mODR error trials. **(B-D)** Mean firing rate response of each population to preferred stimuli during mODR trials with correct (red) and error (black) saccades after the Go command. **(B)** Early storage population. **(C)** Late storage population. **(D)** Response population. In all panels, horizontal bars denote a permutation test over the difference in firing rates across conditions during the last 500 ms before the Go command. “N.S” denotes $p>0.05$, and an asterisk denotes $p<0.05$ (permutation test). All statistical findings hold after decimating trials, confirming they are not attributable to biased sampling or low statistical power.

Figure S6. Summary of LFP Properties. Data show electrodes on which units from only one task selectivity class were recorded. (**A**) Fraction of LFP sites with positive (black) or inverted (red) spatial tuning by frequency for (**i**) early storage, (**ii**) late storage or (**iii**) response unit electrodes. (**B**) Fraction of LFP sites with memory (red) or visual (blue) task-selectivity by frequency for preferred target trials on (**i**) early storage, (**ii**) late storage or (**iii**) response unit electrodes. (**C-E**) Spike-field coherence magnitude versus time for units in each network and fields recorded on a different electrode, at the site of any other delay-tuned unit with the same preferred target location. Coherence estimates were obtained from a 1 s sliding window using 10 Hz bandwidth at frequencies above 13 Hz and 4 Hz bandwidth at lower frequencies. The arrow in each panel indicates the time point shown in Figure 4 using the same analysis parameters. (**C**) Early storage population SFC during trials with (**i**) the fastest 50% and (**ii**) the slowest 50% of reaction times after the Go command. **(D)** Late storage population SFC during (**i**) fast RT and (**ii**) slow RT trials. (**E**) Response population SFC during (**i**) fast RT and (**ii**) slow RT trials.

Supplemental References

1. Funahashi S, Bruce CJ, Goldman-Rakic PS (1993) Dorsolateral prefrontal lesions and oculomotor delayed-response performance: evidence for mnemonic “scotomas.” *J Neurosci* 13(4):1479–1497.

2. Markowitz DA, Wong YT, Gray CM, Pesaran B (2011) Optimizing the decoding of movement goals from local field potentials in Macaque cortex. *J Neurosci* 31(50):18412–18422.

3. Gray CM, Goodell B, Salazar R, Baker J (2006) Semi-chronic recording of neuronal activity in monkey visual cortex using a 60-channel microdrive. *Society for Neuroscience Annual Meeting*, p 481.16.

4. Gray CM, Goodell B (2007) A high-density, large-scale, distributed recording system for semi-chronic monitoring of cortical and sub-cortical neuronal activity in alert monkeys. *Society for Neuroscience Annual Meeting*, p 624.8.

5. Gray CM, Goodell B (2009) A large-scale, distributed recording system for semi-chronic monitoring of cortical and sub-cortical neuronal activity in alert monkeys-iv. *Society for Neuroscience Annual Meeting*, p 390.21.

6. Mante V, Sussillo D, Shenoy K V., Newsome WT (2013) Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* 503(7474):78–84.

7. Zohary E, Shadlen MN, Newsome WT (1994) Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature* 370(6485):140–143.

8. Mitra PP, Pesaran B (1999) Analysis of Dynamic Brain Imaging Data. *Biophys J* 76(2):691–708.

9. Dombrowski SM, Hilgetag CC, Barbas H (2001) Quantitative architecture distinguishes prefrontal cortical systems in the rhesus monkey. *Cereb Cortex* 11(10):975–988.

10. Kantz H, Schreiber T (1997) *Nonlinear time series analysis* (Cambridge University Press, Cambridge, UK).

11. Constantinidis C, Goldman-Rakic PS (2002) Correlated discharges among putative pyramidal neurons and interneurons in the primate prefrontal cortex. *J Neurophysiol* 88(6):3487–3497.

12. Gonzalez-Burgos G, Kroener S, Seamans JK, Lewis D a, Barrionuevo G (2005) Dopaminergic modulation of short-term synaptic plasticity in fast-spiking interneurons of primate dorsolateral prefrontal cortex. *J Neurophysiol* 94(6):4168–77.

13. Diester I, Nieder A (2008) Complementary contributions of prefrontal neuron classes in abstract numerical categorization. *J Neurosci* 28(31):7737–47.

14. Johnston K, DeSouza JFX, Everling S (2009) Monkey prefrontal cortical pyramidal and putative interneurons exhibit differential patterns of activity between prosaccade and antisaccade tasks. *J Neurosci* 29(17):5516–24.

15. Hussar CR, Pasternak T (2009) Flexibility of sensory representations in prefrontal cortex depends on cell type. *Neuron* 64(5):730–43.

16. Crammond DJ, Kalaska JF (1996) Differential relation of discharge in primary motor cortex and premotor cortex to movements versus actively maintained postures during a reaching task. *Exp Brain Res* 108(1):45–61.

17. Zhou X, Katsuki F, Qi X-L, Constantinidis C (2012) Neurons with inverted tuning during the delay periods of working memory tasks in the dorsal prefrontal and posterior parietal cortex. *J Neurophysiol* 108(1):31–8.

18. Pesaran B, Pezaris JS, Sahani M, Mitra PP, Andersen RA (2002) Temporal structure in neuronal activity during working memory in macaque parietal cortex. *Nat Neurosci* 5(8):805–811.

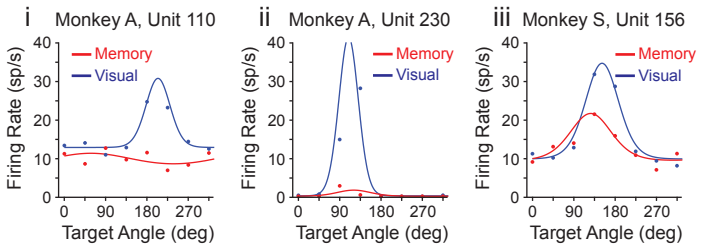
19. Maris E, Schoffelen J-M, Fries P (2007) Nonparametric statistical testing of coherence differences. *J Neurosci Methods* 163(1):161–75.

20. Bruce CJ, Goldberg ME (1985) Primate frontal eye fields. I. Single neurons discharging before saccades. *J Neurophysiol* 53(3):603–35. Available at: http://www.ncbi.nlm.nih.gov/pubmed/3981231.

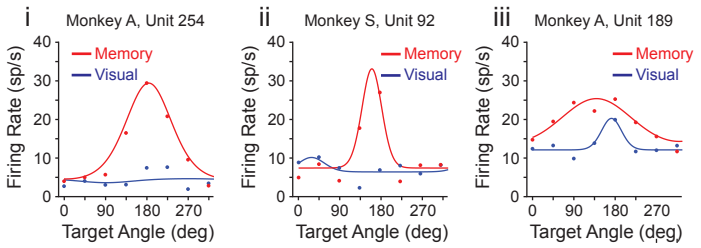
21. Gregoriou GG, Gotts SJ, Desimone R (2012) Cell-Type-Specific Synchronization of Neural Activity in FEF with V4 during Attention. *Neuron* 73(3):581–94.

Figure S1

a Early Storage Unit Late Delay Tuning Curve Examples



b Late Storage Unit Late Delay Tuning Curve Examples



c Response Unit Late Delay Tuning Curve Examples

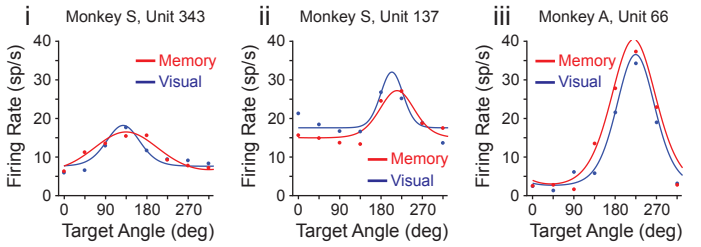


Figure S2

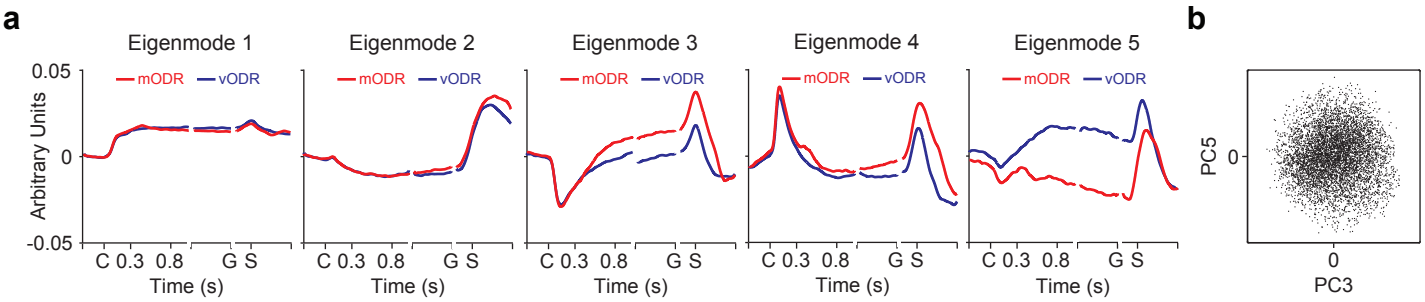

a
Eigenmode 1
Eigenmode 2
Eigenmode 3
Eigenmode 4
Eigenmode 5
mODR
vODR
0.05
0
-0.05
Arbitrary Units
C 0.3 0.8 G S
Time (s)
b
PC5
PC3
0

Figure S3

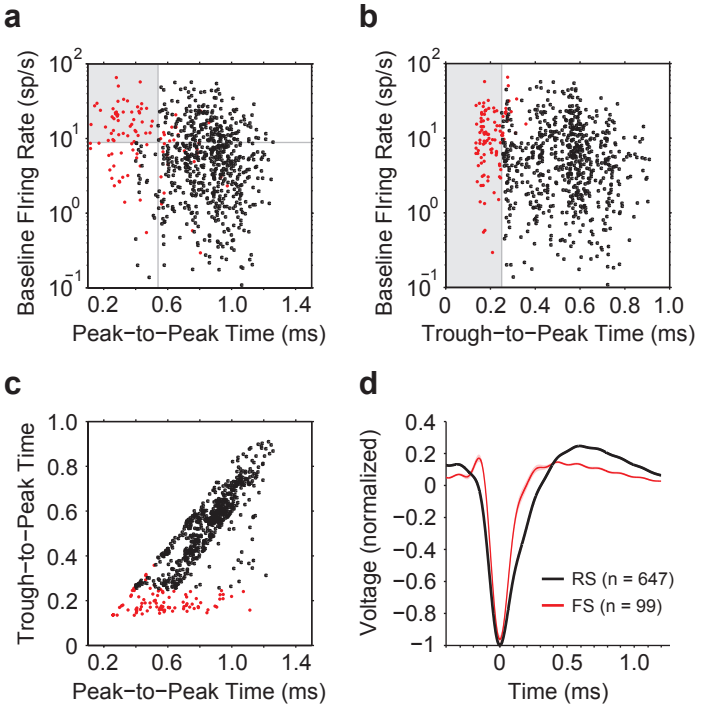

a
b
c
d
Baseline Firing Rate (sp/s)
Peak−to−Peak Time (ms)
Trough−to−Peak Time (ms)
Trough−to−Peak Time
Voltage (normalized)
Time (ms)
RS (n = 647)
FS (n = 99)

Figure S4

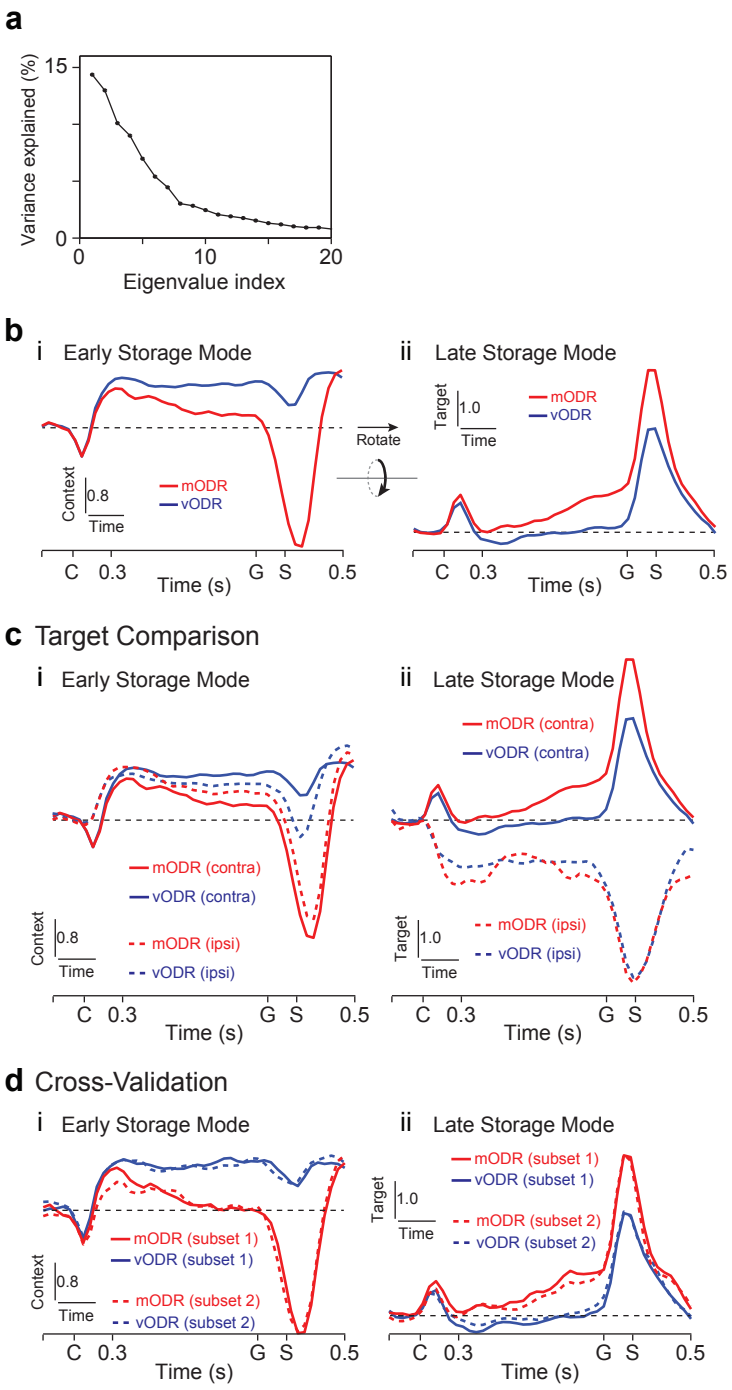
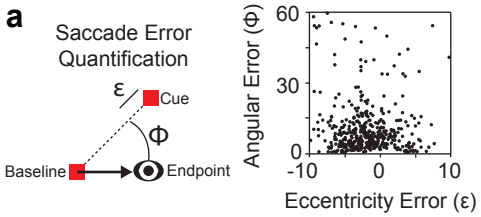
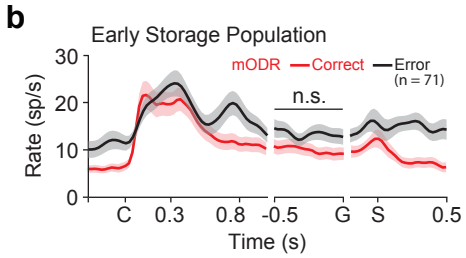
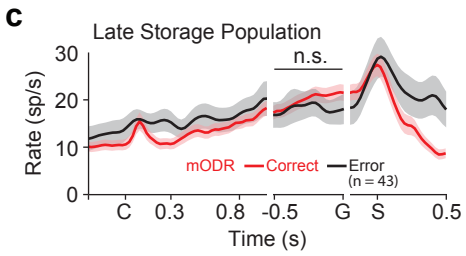

a
Variance explained (%)
15
0
0
10
20
Eigenvalue index
b
i Early Storage Mode
ii Late Storage Mode
Rotate
Context
0.8
Time
mODR
vODR
Target
1.0
C 0.3
Time (s)
G S 0.5
c Target Comparison
i Early Storage Mode
ii Late Storage Mode
mODR (contra)
vODR (contra)
mODR (ipsi)
vODR (ipsi)
d Cross-Validation
i Early Storage Mode
ii Late Storage Mode
mODR (subset 1)
vODR (subset 1)
mODR (subset 2)
vODR (subset 2)

Figure S5


a
Saccade Error Quantification
ε
Cue
Φ
Baseline
Endpoint
Angular Error (Φ)
60
30
0
-10
0
10
Eccentricity Error (ε)


b
Early Storage Population
mODR
Correct
Error
(n = 71)
n.s.
Rate (sp/s)
30
20
10
0
C
0.3
0.8
-0.5
G
S
0.5
Time (s)


c
Late Storage Population
n.s.
Rate (sp/s)
30
20
10
0
mODR
Correct
Error
(n = 43)
C
0.3
0.8
-0.5
G
S
0.5
Time (s)

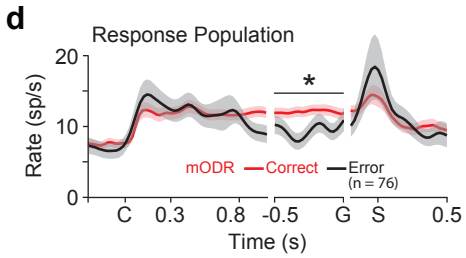
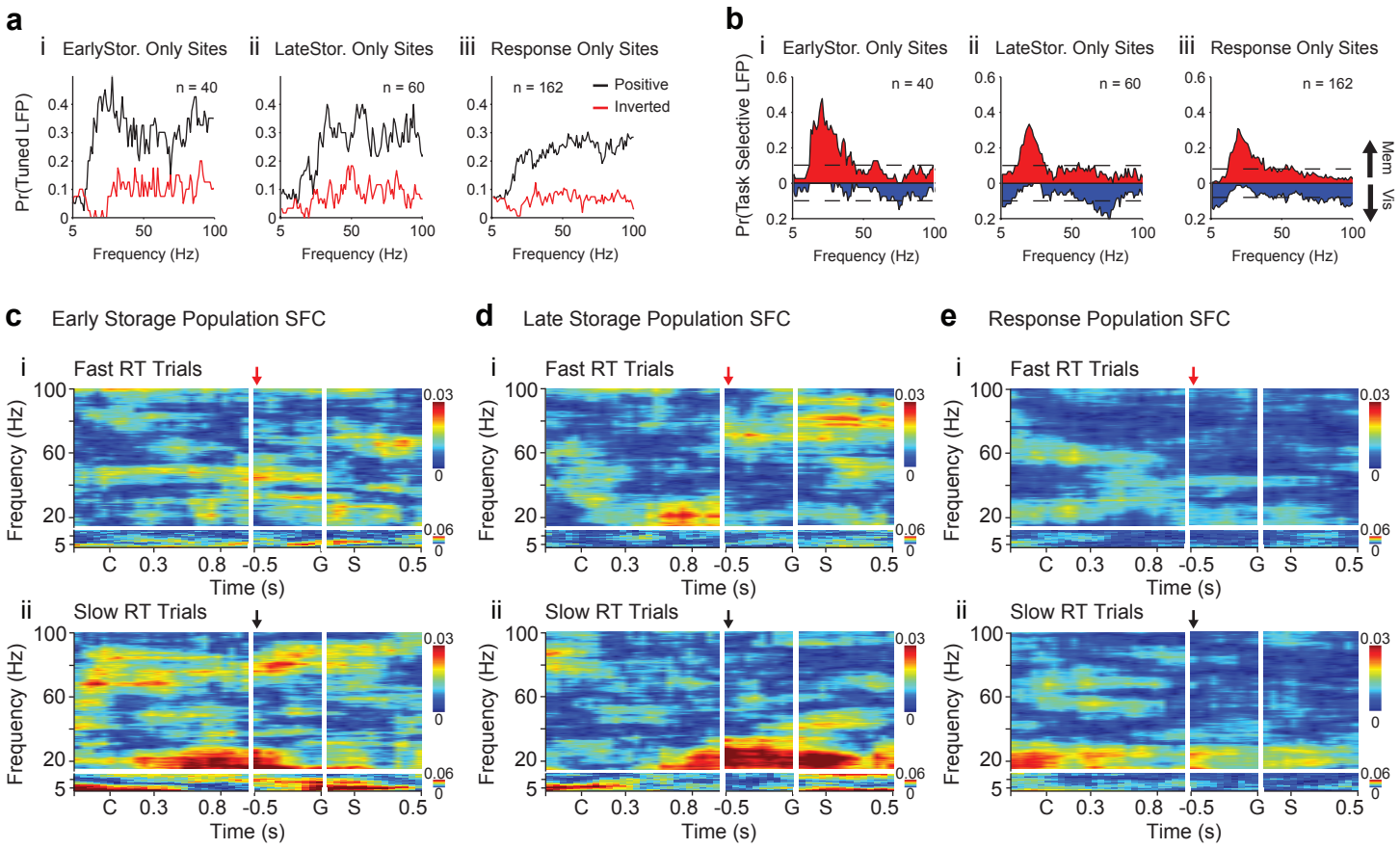

d
Response Population
*
Rate (sp/s)
20
10
0
mODR
Correct
Error
(n = 76)
C
0.3
0.8
-0.5
G
S
0.5
Time (s)

Figure S6


a
i EarlyStor. Only Sites
ii LateStor. Only Sites
iii Response Only Sites
Pr(Tuned LFP)
n = 40
n = 60
n = 162
Positive
Inverted
Frequency (Hz)
b
i EarlyStor. Only Sites
ii LateStor. Only Sites
iii Response Only Sites
Pr(Task Selective LFP)
Mem
Vis
Frequency (Hz)
c Early Storage Population SFC
d Late Storage Population SFC
e Response Population SFC
i Fast RT Trials
ii Slow RT Trials
Frequency (Hz)
Time (s)
C 0.3 0.8 -0.5 G S 0.5