Published in final edited form as:

Neuron. 2018 July 11; 99(1): 215–226.e4. doi:10.1016/j.neuron.2018.05.026.

Feature-Based Visual Short-Term Memory Is Widely Distributed and Hierarchically Organized

Nicholas M. Dotson[1,2,3,*], **Steven J. Hoffman**[1], **Baldwin Goodell**[1], **Charles M. Gray**[1,*]

[1]Department of Cell Biology and Neuroscience, Montana State University, Bozeman, MT 59717, USA

[2]Present address: Department of Bioengineering, University of California, Berkeley, Berkeley, CA 94708, USA

[3]Lead Contact

SUMMARY

Feature-based visual short-term memory is known to engage both sensory and association cortices. However, the extent of the participating circuit and the neural mechanisms underlying memory maintenance is still a matter of vigorous debate. To address these questions, we recorded neuronal activity from 42 cortical areas in monkeys performing a feature-based visual short-term memory task and an interleaved fixation task. We find that task-dependent differences in firing rates are widely distributed throughout the cortex, while stimulus-specific changes in firing rates are more restricted and hierarchically organized. We also show that microsaccades during the memory delay encode the stimuli held in memory and that units modulated by microsaccades are more likely to exhibit stimulus specificity, suggesting that eye movements contribute to visual short-term memory processes. These results support a framework in which most cortical areas, within a modality, contribute to mnemonic representations at timescales that increase along the cortical hierarchy.

In Brief

Dotson et al. recorded from 42 cortical areas in monkeys performing a feature-based memory task. They find that task-dependent differences in firing rates are widely distributed, while stimulus-specific changes in firing rates are more restricted and hierarchically organized.

[*]Correspondence: dotson.neuroscience@gmail.com (N.M.D.), cmgray.montana@gmail.com (C.M.G.).

AUTHOR CONTRIBUTIONS

N.M.D. and C.M.G. designed the experiments, performed the surgical procedures, collected and analyzed the data, and wrote the paper. S.J.H. assisted with data collection and analysis and manuscript preparation. B.G. assisted with all design and fabrication aspects of the project.

DECLARATION OF INTERESTS

The authors declare no competing interests.

SUPPLEMENTAL INFORMATION

Supplemental Information includes five figures and one table and can be found with this article online at https://doi.org/10.1016/j.neuron.2018.05.026.

INTRODUCTION

Working memory is essential to cognition. It enables the short-term retention and utilization of behaviorally relevant information for virtually all cognitive tasks (Baddeley, 2003). The neural mechanisms that mediate feature-based visual working memory have been intensively studied for several decades (Luck and Vogel, 2013; Sreenivasan et al., 2014; Lara and Wallis, 2015; Christophel et al., 2017). In non-human primates (NHPs), feature-based mnemonic representations (measured as differences in neural firing rates between stimuli held in memory) have been demonstrated in prefrontal (Fuster et al., 1982; Quintana et al., 1988; Miller et al., 1996; Peng et al., 2008; Salazar et al., 2012; Mendoza-Halliday et al., 2014), posterior parietal (Sereno and Maunsell, 1998; Salazar et al., 2012; Sarma et al., 2016), inferotemporal (Fuster and Jervey, 1981; Miyashita and Chang, 1988; Miller et al., 1991, 1993), and extra-striate visual cortical areas (Bisley et al., 2004; Mendoza-Halliday et al., 2014). Functional imaging studies have corroborated and extended these findings in humans (Courtney et al., 1997; Owen et al., 1998; D’Esposito et al., 2000; Postle et al., 2003).

However, because our understanding of feature-based visual working memory is derived from studies using different tasks, stimuli, and recording techniques, the full extent of the participating circuit and the underlying neural mechanisms remains highly debated. In particular, fMRI-based decoding analyses of the content of visual short-term memory (Harrison and Tong, 2009; Serences et al., 2009; Ester et al., 2015) have revealed involvement of early visual cortex, prompting debate on the nature and role of sensory areas in the storage and maintenance of mnemonic representations. A salient issue is whether feature-based mnemonic representations are present in neural spiking activity in early visual areas (Serences 2016). However, the few studies that have made such measurements have reached conflicting conclusions (Bisley et al., 2004; Zaksas and Pasternak, 2006; Mendoza-Halliday et al., 2014). The role of prefrontal cortex is also heavily debated. Many argue that it is primarily involved in executive functions (Sreenivasan et al., 2014; D’Esposito and Postle, 2015; Lara and Wallis, 2015), while others emphasize a key role in both executive functions and mnemonic representations (Serences, 2016; Hasson et al., 2015; Christophel et al., 2017). Finally, because studies of neural spiking activity typically focus on one cortical area at a time, the relative contribution of widely distributed cortical areas to working memory processes is largely unknown. Resolving these issues is necessary for developing a mechanistic theory of how and where working memory processes are carried out.

Here, we focus on two basic questions about feature-based visual short-term memory that remain largely unanswered. Which cortical areas are involved in the neural circuit that mediates visual short-term memory, and how do the individual components of this circuit differ in their functional roles? To address these questions, we performed large-scale microelectrode recordings of neuronal activity in NHPs performing a feature-based visual short-term memory task and determined both the task dependence and stimulus selectivity of the recorded neurons.

RESULTS

We recorded broadband neuronal activity from a total of 42 cortical areas in two NHPs (monkeys E and L) while they performed a feature-based delayed match-to-sample (dMTS) task and an interleaved visual fixation task. The dMTS task required the monkeys to remember a centrally presented sample image (1 of 5 possible images) for a minimum of 800 ms (800–1,200 ms in monkey E; 1,000–1,500 ms in monkey L) before making a choice between a matching and a non-matching image (Figure 1A). During the fixation task, occurring on ~10% of the trials, the monkeys simply had to fixate the central target for the same duration as the dMTS task. Data were collected using large-scale semi-chronic recording devices with up to 256 independently moveable microelectrodes (Dotson et al., 2015, 2017). Neuronal spiking activity was extracted and sorted offline. Simultaneous recordings were made from up to 12 and 29 different cortical areas in monkeys E and L, respectively. The recording locations and sample sizes, combined from both animals, are shown in Figure 1B (Table S1). Sparsely sampled cortical areas, and areas with similar functional properties (e.g., somatosensory areas 1, 2, and 3) were merged with adjacent areas, resulting in a total of 24 different areas/groups combined across monkeys (Tables 1 and S1). Figure 1C shows an example of eye position signals (bottom) and broadband electro-physiological data sampled simultaneously from microelectrodes located in 27 separate cortical areas in monkey L.

Prior to determining the task dependence and stimulus selectivity of each cortical area, we first identified units that showed modulations in their firing rates locked to microsaccades (Martinez-Conde et al., 2013; STAR Methods; Table S1; Figures S1 and S2). These microsaccade-modulated (MSM) units occurred in 22 of the 24 cortical areas/groups in our sample but were most common (>10% incidence) in areas V1, V2, DP, and 8L (Figure S2A). Because of the possible confound introduced by eye movement related activity, we excluded all of these MSM units from our analyses unless otherwise stated. Second, because our recordings in areas V1 and V2 spanned a large portion of the retinotopic map, many of the units in these areas had receptive field locations (i.e., >2° eccentricity) that prevented them from responding directly to the sample stimuli. We therefore identified units in areas V1 and V2 that displayed a short-latency (40–100 ms) excitatory visual response (SLVR) to at least one of the sample stimuli, and we analyzed these units separately from the remaining units in V1 and V2 (STAR Methods; Table S1).

Task-Dependent Activity Is Widely Distributed

The first objective of our analysis was to determine the extent to which each unit differentially participated in the two tasks. We assessed this “task dependence” by comparing the firing rates during the dMTS task (combined across all 5 stimuli) to the firing rates occurring during the interleaved fixation task, using a 200-ms time bin, stepped every 50 ms (Wilcoxon rank-sum test, $p < 0.05$, FDR [false discovery rate] corrected). As mentioned above, all MSM units were excluded from this analysis. Figure 2A (I-XII) shows example peri-stimulus time histograms (PSTHs) from units in 9 different cortical areas sampled from both monkeys. Each plot shows the average firing rate for the dMTS (blue) and fixation (red) tasks for an individual unit from the cortical area indicated at the top (significant bins are marked with a black square). In general, the neuronal responses during the two tasks differed substantially, but often in unexpected and heterogeneous ways. For example, the selected units from 9/46V, V2, and V1 (I, X, and XII) conformed to our expectation of a relatively stable rate during the fixation trials and a clear task-dependent modulation of rate during the dMTS task. However, we also found instances of the opposite pattern, where units displayed a stable firing rate during the dMTS task and a robust rate modulation during the fixation task (V [F1] and VI [3]). Moreover, we found many instances of a more complex multi-phasic relationship between the two tasks. These included cases ranging from a simple push-pull pattern, where a biphasic change in firing rate during one of the tasks was mirrored by a similar but opposite pattern during the other task (II [46v], IV [F7], and VII [anterior intraparietal area; AIP]) to cases where the rates would diverge early in the tasks and converge back to a similar rate by the end of the tasks (VIII [AIP], IX [V2], XI [V1]). Interestingly, we also found numerous instances of ascending rates during the fixation task in areas of the visual hierarchy as early as V1 (XI). These and many other examples demonstrated that task-dependent changes in rate are widely and heterogeneously distributed across the cortex and that activity during the fixation task is itself highly dynamic, indicating that this task reflects a distinct cognitive process.

To determine the task dependence for each cortical area/group, we first calculated the incidence of significant differences in firing rates between the two tasks. We then separated the resulting distribution at each time point according to whether each unit’s response during the dMTS task was greater than (enhanced) or less than (suppressed) the response during the fixation task. The plots in Figures 2B and 2C show the results of these calculations for ventral prefrontal cortex (vPFC) ($n = 124$ units). The incidence of task-dependent differences in activity increased rapidly during the sample period to a peak value near 50% and remained near 40% throughout the delay period (Figure 2B). The colored bar at the top shows the same data plotted as a heatmap. Figure 2C shows that the incidence of enhancement and suppression relative to the fixation task is split roughly equally throughout the task for vPFC. To visualize the ratio of enhancement to suppression, we calculated a rate modulation index (RMI) as a function of time (shown as a heatmap in Figure 2C). This is the percentage difference of enhanced (E) and suppressed (S) units at each time bin ($\mathrm{RMI} = [(\#E - \#S)/(\#E + \#S)] \times 100$). To determine if the incidences of these two processes differed from one another, we tested the two counts at each time bin using a binomial test ($p < 0.05$, FDR corrected). In this cortical area (vPFC), we found no significant differences and conclude that there is an equal distribution of activity during the dMTS task where the firing rates are greater or less than the rates occurring during the interleaved fixation task. We refer to this as balanced activity.

We applied these calculations to all 24 cortical areas/groups (Figure S3). The incidence of significant differences in firing rates revealed widespread differential involvement in the two tasks (Figure 3A). Every areal group we studied demonstrated some degree of task dependence that could broadly be separated into three types of activity profiles: visually responsive, ramping, and sustained. Visually responsive areas showed clear sample related task dependence, which either tapered off gradually (e.g., V2-SLVR) or remained elevated (e.g., vPFC and AIP/VIP [ventral intraparietal area]) throughout the delay period. In several other areas, the incidence of task dependence began ramping up in the middle of the sample period and continued until the end of the delay (e.g., F2 and F7). Finally, many of the areas simply showed weak but sustained task-dependent activity during all task epochs (e.g., dorsal prefrontal cortex [dPFC] and 7b). Interestingly, primary motor (F1) and somatosensory (1/2/3) areas were also clearly task dependent.

Analysis of the differences in firing rates between the two tasks (enhancement or suppression) revealed a more complex pattern of the task dependence across the cortex. To visualize this pattern, we plotted the RMI as a function of time for all 24 areas/groups (Figure 3B). We then tested the two counts at each time bin using a binomial test ($p < 0.05$, FDR corrected) and marked each significant bin with a white cross. As expected, we found that SLVR units in V1 and V2 were strongly enhanced during the sample period but then were suppressed during the late phase of the fixed delay and returned to a balanced state prior to the match onset (bottom of Figure 3B). The large majority of units in V1 and V2 (i.e., those having receptive field locations >2° eccentricity) were suppressed during both the sample and delay periods, and displayed a rebound enhancement during the match-locked period. A number of other cortical areas also displayed suppression throughout much of the task, including the primary somatosensory areas (1/2/3) and medial posterior parietal areas (7op and 5/MIP [medial intraparietal area]). The remaining cortical areas, including most prefrontal (e.g., vPFC and dPFC) and lateral posterior parietal areas (e.g., AIP/VIP and 7b), exhibited a balanced distribution of enhancement and suppression. To further visualize how these relationships are distributed across the cortex, we plotted the occurrence of enhanced, suppressed, and balanced activity during the end of the fixed delay period (gray shading and arrows in Figure 3B) on a cortical flatmap (Figure 3C). This revealed an interesting pattern during much of the delay period, where early sensory areas are suppressed and the remaining areas are primarily balanced. Thus, while the responses in association areas (e.g., vPFC) are highly dynamic and heterogeneous between the two tasks, they tend to balance out at the population level. Early sensory areas, on the other hand (i.e., V1 and V2, and somatosensory areas 1, 2, and 3), tend to be suppressed at the population level during the delay.

Embedded Hierarchy for Mnemonic Representations

The previous analysis revealed that task-dependent activity is widely distributed. Do the same areas encode the stimulus identity? To address this question, we determined the time course of stimulus-selectivity during the dMTS task by comparing the firing rates across stimuli in 200-ms windows with a 50-ms step (Kruskal-Wallis test, $p < 0.05$, FDR corrected). As with the previous analysis, all MSM units were excluded from this analysis. Figure 4A (I-XIV) shows example PSTHs with stimulus-selective responses from units recorded in 11 different cortical areas, revealing a striking variety of stimulus-specific responses among both association and sensory areas. Example units in vPFC (I and II) and posterior parietal cortex (VI–IX) displayed highly selective activity in response to the different stimuli that lasts throughout the sample and delay periods. Other units in prefrontal areas exhibit transient selectivity that is superimposed on a background of apparent suppression (III) or that ramps up during the delay period (IV–V). Activity in early visual areas V1, V2, and V4 (X–XIV) also displayed transient and sustained periods of selectivity. This occurred even when the delay-period firing rates were quite low (XI) or ramping up during the delay (XII).

To characterize the overall behavior of the data, we calculated the incidence of selectivity as a function of time for each of the 24 cortical areas/groups. The result of this analysis for area vPFC is shown in Figure 4B. And the results for all areas/groups are shown in Figure 5A (also see Figure S4 for line plots of these data). This analysis revealed a pattern of stimulus selectivity that was sparser compared to the widespread distribution of task dependence (Figure 3). Units in V1 and V2 that were excited by the sample stimuli (bottom two plots in Figure 5A, SLVR) displayed a high incidence of stimulus selectivity throughout the sample and well into delay. This delay period selectivity occurred even though the firing rates during this period were typically low and suppressed relative to the fixation task (Figures 3B and 3C). Other cortical areas displayed more sustained stimulus selectivity throughout the delay (e.g., vPFC, 8L, AIP/VIP, and 7b) (Figure 5A), and a number of areas showed little or no selectivity throughout the task (e.g., F1, 7op, 5/MIP, 1/2/3, and V6A) (Figure 5A). To further characterize these effects, we calculated the median incidence of selectivity during the late delay period for all areas/groups (Figure 5A, gray shading and arrows). This revealed that areas V1-SLVR, V2-SLVR, V4, LIP, 7b, AIP/VIP, F7, 8L, and vPFC have significant delay period selectivity that exceeds an incidence of 5% (Figure 5B).

These results suggest that under the conditions of this experiment, a subset of areas contribute to short-term memory maintenance. However, because it was impractical to adjust the position and properties of the sample stimuli to optimally activate each unit simultaneously, differences in the absolute incidence of delay period selectivity may be less informative than expected. Therefore, we chose to analyze the incidence of late delay period selectivity, relative to that occurring throughout the task, in order to determine the contribution of each area to short-term memory maintenance. This measure, illustrated in Figure 6A, reveals that the relative incidence of stimulus selectivity during the late delay period is substantially higher in vPFC than V2-SLVR, even though the absolute incidence of selectivity during this period is marginally higher in V2-SLVR (Figure 5B). To see how this measure varies across the cortex, we plotted the relative incidence of delay period selectivity in rank order for those areas/groups with an absolute incidence of delay period selectivity exceeding 5% (Figure 6B). This revealed that the relative incidence of stimulus selectivity decays rapidly in early visual areas, while it is more sustained in prefrontal and posterior parietal areas, thereby forming a functional short-term memory hierarchy. The inset in Figure 6B shows an anatomically derived hierarchy from the studies by Markov et al. (2014b) and Chaudhuri et al. (2015) for areas that match or are included in the significant areas/groups. The functional hierarchy agrees well with the anatomical hierarchy. Finally, within this functional hierarchy, we find a high incidence of units with responses that are both stimulus selective and task dependent throughout the task (Figure S5).

These analyses indicate that the incidence of stimulus selectivity tends to peak during the sample period and then differentially decays during the delay in a manner that reflects the anatomical hierarchy. We suspected that the latter effect might also be due in part to the recruitment of newly stimulus-selective units during the delay period in areas where this decay is less pronounced (see examples in Figure 4A, III-V and X). To address this question, we identified the first time bin that a unit became stimulus selective. For instance, if a unit responds selectively to the sample stimulus and then remains selective throughout the task, then we only count the first time bin during the sample period that the unit is selective and discard the other time bins from the analysis. This provided us with histograms of the newly recruited stimulus selective units. We normalized these histograms and computed the cumulative sum over time for the areas showing >5% significant delay period selectivity (Figure 6C). The time when the cumulative sum reaches 1 indicates how long into the dMTS task selective units continued to be recruited. In Figure 6C, we see that early visual areas peak almost immediately after the sample onset, indicating that no units are recruited during the delay that were not already selective during the sample period (e.g., V1-SLVR and V2-SLVR). Areas higher in the hierarchy contain units that become stimulus selective later in the task throughout the delay period. Figure 6D shows the rank ordered time to the last recruited units (cumulative sum = 1). The results from this analysis match the general hierarchical scheme derived from the relative incidence of delay-period selectivity (Figure 6B), with V2-SLVR and V1-SLVR low in the hierarchy; LIP, 8L, and V4 in the middle; and 7b, vPFC, F7, and AIP/VIP at the highest level.

Microsaccades Encode Visual Memories

These results demonstrate that widespread cortical areas involved in visual processing and perception contribute to short-term mnemonic representations. What is the relationship between perception and memory? Specifically, are mnemonic representations entirely abstract, or do they maintain a semblance of the real image (i.e., a shape)? Because the animals routinely made microsaccades (typically <1 degree of visual angle [dva]) to different portions of the sample images (Figure 1C, bottom plots), we posited that if the mnemonic representations maintain a spatial form, then the eye position during the delay period would encode the remembered image. Figure 7A shows an example of the microsaccade endpoints for two stimuli during a recording session in Monkey L. In the presample period, the microsaccade endpoints are overlapping, while during the sample and delay periods, they are largely nonoverlapping, supporting our hypothesis. To test for this, we used mutual information analysis of the microsaccade endpoints for the same session. This analysis revealed that stimulus-specific microsaccades occur during the sample and late delay periods of the task (Figure 7B). The overall incidence of this effect across all recording sessions was similar for both monkeys (Figure 7C), demonstrating that eye position following microsaccades reliably encodes mnemonic representations during the end of the fixed delay period. This suggests that the monkeys were scrutinizing their short-term visual memories of the sample stimuli.

These findings, along with the well-established relationship between microsaccades and neural activity (Martinez-Conde et al., 2013), further imply that microsaccadic eye movements may be an integral part of visual short-term memory. Since we found that microsaccades modulate activity in a large number of the units in our sample (Figures S1 and S2; STAR Methods), we sought to determine if these units contribute differentially to stimulus selectivity during the task. (It is important to note that all significant MSM units were excluded from all of the previous analyses.) To accomplish this, we compared the incidence of stimulus selectivity between units that were and those that were not modulated by microsaccadic eye movements (Pearson's chi-square test of independence, $p < 0.05$, FDR corrected). We restricted our analysis to cortical areas V1, V2, and 8L, which contained a minimum of 25 MSM units (Table S1). We did not analyze the MSM SLVR units in V1 and V2 because of low sample sizes (e.g., 10/98 in V1 and 9/107 in V2). This analysis revealed a higher incidence of stimulus selectivity in the MSM units (Figures 7D-7F). In V1 and 8L, this effect occurs in the sample period and throughout the fixed delay. In V2, the differences occur primarily around the sample offset. These results indicate that MSM units tend to have a higher incidence of stimulus selectivity that extends into the delay period of the task.

Finally, given the higher incidence of stimulus selectivity in MSM units, we reran the stimulus-selectivity analyses with MSM units included in order to determine if this had an effect on the functional hierarchy. We found no significant differences in the overall incidence of delay-period selectivity and no change in the functional hierarchy as described in Figures 6B and 6D. This is likely because the MSM units only compose a small fraction of the total units.

DISCUSSION

To elucidate the neural circuit dynamics underlying visual short-term memory, we developed a large-scale microelectrode recording device that encompasses an entire cerebral hemisphere (Dotson et al., 2017) and analyzed recordings from a total of 42 cortical areas in two NHPs performing a feature-based dMTS task and an interleaved visual fixation task. We find that the cortical circuit defined by the differences in activity between these two tasks (task dependence) is widely distributed, heterogeneous, and dominated by population activity during the dMTS task that is either balanced or suppressed relative to the fixation task. Thus, even a simple cognitive task, such as remembering an item during a brief delay, recruits widespread changes in activity throughout the cortex. Embedded within this large-scale circuit, we identified a functional hierarchy for mnemonic representations. The hierarchy is expressed as an increase in the relative incidence of stimulus selectivity during the delay, and an increase in the latency in which newly selective units are recruited. This hierarchy extends from the early visual cortex to high-level association areas in prefrontal and posterior parietal regions, and it closely matches the hierarchy derived from anatomical measurements of feed-forward and feedback connections (Markov et al., 2014b; Chaudhuri et al., 2015). Stimulus-selective activity within this hierarchy occurs in conjunction with heterogeneous task-related information (Figure S5). This apparent mixed selectivity may result in high dimensional encoding of all stimulus and task information at each hierarchical level, similar to what has been observed in prefrontal cortex (Rigotti et. al., 2013). It may also endow these areas with varying degrees of distractor resistance, depending on the concentration and types of mixed selective units (Parthasarathy et al., 2017). We also identified a behavioral correlate of visual short-term memory. Microsaccadic eye movements during the memory delay encode the stimuli held in memory, and units modulated by microsaccades are more likely to exhibit stimulus-specific activity.

Methodological Caveats

While there are clear advantages to performing large-scale simultaneous recordings of neural activity, our approach also introduced several experimental limitations that likely influenced our findings. The first concerns the sampling of activity. We collected less data in monkey E than monkey L because of an improvement in the design and implementation of the recording methods during the time spanning the two experiments (see Dotson et al., 2017). Consequently, our findings for some cortical areas are based on data from one, but not both, animals. Similarly, the sample sizes for some cortical areas required us to combine data among adjacent areas in order to validate our statistical tests. This is not uncommon in some physiological studies of cortex, but it does reduce the specificity of some of our findings. Our experimental design also prevented us from performing population decoding analyses (e.g., Mendoza-Halliday et al., 2014; Parthasarathy et al., 2017), which typically rely on combining data across recording sessions that use the same stimuli or high-density recordings of individual cortical areas.

We also were unable to reliably measure neural activity from the ventral temporal visual pathway, specifically areas TEO and TE. This was due to limitations in the design of the device implanted on monkey E and some failures in the actuator mechanisms in the device used in monkey L. Therefore, we were unable to characterize neural activity in a major division of the cortical pathway underlying feature-based vision and short-term memory.

Additionally, because we sampled neural activity from many different cortical areas simultaneously, it became impractical to tailor the parameters of the experiment (e.g., the location and properties of the stimuli) to each recording site. Thus, many of the recorded neurons may have been unresponsive or weakly responsive to the stimuli and parameters of the task. This likely led us to underestimate the incidence and magnitude of task dependence and stimulus selectivity. However, our method also enabled us to obtain an unbiased estimate of the distribution of task-dependent and stimulus-selective activity over widespread areas of cortex that otherwise may have gone undetected.

Memory Maintenance in V1 and V2

Our findings help to resolve an ongoing debate regarding the contribution of early visual cortical areas to feature-based visual short-term memory. Functional imaging studies in humans have repeatedly demonstrated the ability to decode short-term memory content from primary and early extrastriate areas of visual cortex (Harrison and Tong, 2009; Serences et al., 2009), even though elevated delay-period activity is largely absent in these areas (Riggall and Postle, 2012). These findings support the concept that mnemonic representations in these areas may be mediated by top-down input (Mendoza-Halliday et al., 2014) and/or changes in synaptic strength induced by the sample stimulus (see Serences, 2016 for review). However, studies investigating memory-related unit activity in early areas of visual cortex have reached somewhat conflicting conclusions with respect to this hypothesis (Supèr et. al., 2001; Bisley et al., 2004; Lee et al., 2005; Zaksas and Pasternak, 2006; Mendoza-Halliday et al., 2014; van Kerkoerle et. al., 2017). We find evidence of stimulus-specific spiking activity during feature-based visual short-term memory in V1 and V2. These effects occur at firing rates that are near or below the level of activity measured during the interleaved fixation task (Figures 2 and 3), arguing against the need to postulate a sub-threshold storage mechanism.

Decoding Items in Memory with Microsaccades

Our findings demonstrate that microsaccades provide a behavioral readout of the stimuli held in memory. These results suggest a tight link between perception and short-term memory maintenance. Mnemonic representations may maintain spatial relationships similar to the perceived images, enabling the monkeys to scrutinize these representations. Similar results have been observed in human subjects imagining a previously seen image (Brandt and Stark, 1997; Laeng and Teodorescu, 2002). This oculomotor behavior may result in a sensory-motor feedback loop that facilitates the maintenance of visual working memories. This is supported by our finding of a higher incidence of stimulus selectivity in MSM units in areas V1, V2, and 8L. Interestingly, we find this activity among units in V1 and V2 with receptive fields that likely do not overlap with the sample image (i.e., the non-SLVR units). This may help explain why decoding in functional imaging studies can be done in parts of visual cortex outside of where the sample was presented and even in contralateral areas of cortex (Ester et al., 2009). However, because of our experimental design, we were unable to completely separate the role of delay period selectivity from the activity evoked by microsaccadic eye movements. While it is possible that microsaccades are an integral component of short-term memory, they may also operate in parallel and exert little or no influence on mnemonic representations. Future studies that dissociate these roles will be necessary to determine the full relationship between eye movements and short-term memory.

The Balanced Activity State and Short-Term Memory

A prominent result in the task-dependence analysis is that the incidence of enhanced and suppressed activity is typically balanced in association areas and suppressed in sensory areas during the delay period. This may be linked to the neural mechanisms underlying stimulus-specific persistent activity. We see that a fundamental difference arises at either end of the hierarchy. In early visual areas, the relative incidence of stimulus-specific unit activity decays rapidly and selective unit recruitment ceases after the sample presentation, while high-level association areas maintain a higher relative incidence of stimulus selectivity and continue to recruit selective units well into the delay period. These differences match the pattern of balanced and suppressed task-dependent activity. Collectively, these findings suggest that the balance of enhancement and suppression that occurs during the dMTS task, relative to the interleaved fixation task, is a signature of higher-order cortical areas and may contribute to the maintenance of mnemonic representations and short-term memory in general.

Short-Term Memories Are Maintained in a Hierarchy of Cortical Areas

Our analysis of the relative incidence and onset of delay period selectivity enabled us to identify a functional hierarchy for mnemonic representations. What mechanisms produce the functional hierarchy? Functional imaging (Honey et al., 2012), neurophysiological (Murray et al., 2014), and modeling studies (Chaudhuri et al., 2015) have all identified intrinsic timescales with a hierarchical ordering. One interpretation of these findings is that they provide a hierarchy of temporal receptive windows of increasing size that enable the accumulation and integration of information over increasing periods of time (Hasson et al., 2015). Our findings support this interpretation and are consistent with a framework in which most cortical areas, within a modality, contribute to mnemonic representations at timescales that increase in a hierarchical manner. This framework provides a parsimonious explanation for any cognitive task that requires information to be gathered, combined, and remembered.

STAR★METHODS

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Nicholas M. Dotson (dotson.neuroscience@gmail.com).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Subjects—Data was collected from two female macaque monkeys (Monkey E and Monkey L) while microelectrode recordings were performed using a large-scale microdrive system (Dotson et al., 2017). Further details of the recording technique are provided below. All procedures were performed in accordance with NIH guidelines and the Institutional Animal Care and Use Committee of Montana State University.

METHOD DETAILS

Behavioral task—The monkeys were seated in a primate chair (Gray Matter Research, LLC), head fixed using a cranial head post (Gray Matter Research, LLC), and positioned 57 cm from a 19-inch monitor. MonkeyLogic software was used to run the experiment and record eye position data (Asaad and Eskandar, 2008a, 2008b). Eye position data was acquired using an infrared eye-tracking system (240 Hz; ISCAN) and converted to degrees of visual angle (dva) in MonkeyLogic. The monkeys were trained to perform a feature-based delayed match-to-sample task (Figure 1A). A trial begins when the monkey acquires and holds fixation on a small fixation spot (fixation window diameter = 3 dva). At a latency of 500 ms for Monkey E or 800 ms for Monkey L, one of five possible sample images (size: 2.4x2.4 dva) is presented for 500 ms in the center of the screen (obscuring the fixation point). During the sample period the monkey has to maintain it’s gaze in the same 3 dva window used during the fixation period. Sample images were pseudo-randomly chosen from a pool of 40 or 100 images for Monkey E or Monkey L, respectively. The animals were familiar with all images. The sample stimulus is followed by a randomized delay, 800-1200 ms for Monkey E and 1000-1500 ms for Monkey L, in which no stimulus is present. During this time the monkeys must maintain gaze on the central fixation point. At the end of the delay period, the fixation target is extinguished and the matching image and a non-matching image (one of the four other images chosen randomly) appear 5 dva from the center of the screen. For Monkey E, the match and non-match were always placed across from each other on the horizontal plane. The location (left or right) of the match and non-match were randomized on each trial. For Monkey L, the images were aligned either vertically, horizontally or diagonally. The location of the match and non-match and the alignment was randomly chosen on each trial. While the match image is visible, the monkey must make a saccadic eye movement to the matching image and maintain fixation for a brief period of time (200 ms for Monkey E and 500 ms for Monkey L). Correct trials were rewarded with a drop of juice. Approximately 10% of the trials did not include stimuli and the monkey simply had to maintain visual fixation on the central spot throughout the trial to receive a reward. In order to easily compare these trials with the match-to-sample trials, we used the same timing structure, except that there was no sample image presented and no match period. The animals completed > 500 correct trials on each session, with behavioral performance typically > 75%. Data used in this report are from 25 and 62 recording sessions from Monkey E and Monkey L, respectively.

Recording techniques—Each animal was implanted with a custom made large-scale recording system containing independently movable microelectrodes spanning the length and width of an entire hemisphere (Dotson et al., 2017). Each system consists of a form-fitting recording chamber and a microdrive composed of a guide array, an actuator block, a printed circuit board (PCB), and a screw guide. Linear actuators (256 for Monkey E, 252 for Monkey L) are housed in the actuator block, with a separation of 2.34 mm. Each actuator consists of a miniature stainless steel lead screw, a threaded brass shuttle, and a compression spring. For Monkey E, each actuator provided 20 mm of microelectrode travel at a resolution of 8 turns/mm. For Monkey L, actuators had 33 mm or 41 mm of microelectrode travel at a resolution of 5 turns/mm. The recording systems remained on the animals throughout the entire experiment.

Once the monkeys reached criterion performance on the tasks, we carried out a multi-step implantation sequence and began neural recordings when the animals were fully recovered, healthy and performing the task normally. We gradually moved all of the functioning microelectrodes through the dura and into the cortex over a period of 2-4 weeks. This was done in an incremental manner by advancing a subset of 10-30 microelectrodes each day until unit activity was first detected. Once the recording phase of the experiment began, we made small incremental advancements (~50-500 μm) to a varying subset of electrodes on each recording session. We routinely measured microelectrode impedance and ceased advancing a microelectrode whenever its impedance was < ~50 kΩ or > ~2 MΩ. We attempted to adjust the high impedance microelectrodes and recover the signal. If this failed, we considered the actuator or microelectrode to be damaged and did not move these microelectrodes further.

We carried out daily recording sessions 3-5 days/week over a period of ~6 and ~9 months, in monkeys E and L, respectively. Broadband neuronal activity was recorded simultaneously from all viable microelectrodes (0.1 Hz - 9 kHz, sampled at 32 kHz) using a Digital Lynx SX recording system (Neuralynx). The reference and ground connections were tied together and connected to the chamber. This created a distributed reference signal.

Histology—Recording locations were determined by combining records of the microelectrode depths and histological information. When recordings were completed, small electrolytic lesions were made at the tip of all functioning microelectrodes (10 μA DC for 25 s). The animals were then euthanized (Pentobarbital, 100mg/kg; i.v.) and perfused through the heart with phosphate buffered saline (PBS) followed by a solution of 5% paraformaldehyde in PBS. After perfusing the animals, we removed the recording systems without retracting the microelectrodes. We used slightly different procedures for Monkey E and Monkey L to perform the reconstructions.

For Monkey E, following perfusion, the brain was removed, and sunk in a solution of fixative with 30% sucrose several days before being sectioned (60 μm) and stained for Nissl substance (FD Neurotechnologies). To reconstruct the brain, the stained sections were photographed and then imported into Free-D (Andrey and Maurin, 2005). Sections were manually registered and then electrolytic lesions and microelectrode tracks were marked on the images. This information provided a 3D reconstruction of all the microelectrode tracks. We used this information and the record of microelectrode depths to estimate the microelectrode tip position and identify the recording locations of each microelectrode on each recording session.

For Monkey L, following perfusion, we removed the top of the skull and then cut the brain in half in the coronal plane at Bregma −15 mm. This ensured that during the sectioning process each slice was in the coronal plane and gave us a crude estimate of the anterior-posterior position of each slice. Each half of the brain was then sunk in a solution of fixative with 30% sucrose for several days before being sectioned (60 μm) and stained for Nissl substance (FD Neurotechnologies). During sectioning, we photographed the frozen block-face of the brain in order to preserve the shape of each slice. Stained sections were used to identify electrolytic lesions and microelectrode tracks. The block-face photographs were imported into the Computerized Anatomical Reconstruction and Editing Toolkit (Caret) (Van Essen, 2012). Each photograph was traced and then annotated with information about the location of electrolytic lesions and microelectrode tracks from the Nissl-stained slices. Then, as with Monkey E, we used the 3D reconstruction and the record of microelectrode depths to estimate the microelectrode tip position and identify the recording locations of each microelectrode on each recording session. For both monkeys, we identified the cortical area or subcortical nucleus for each recording site by comparing the histological reconstructions to the atlas published by Markov et al. (2014a). The flat-maps in Figures 1 and 2 were created using the Scalable Brain Atlas website (Van Essen, 2012).

Anatomical hierarchy—The anatomical hierarchy in Figure 6 was derived from Markov et al. (2014b) and Chaudhuri et al. (2015). Areas V1, V2, V4, 8L, 9/46V, 7b, and F7 were ordered based on results reported by Chaudhuri et al. (2015). Area LIP was placed above 8L based on the study by Markov et al. (2014b).

Spike sorting—The technique for spike sorting follows earlier studies (Salazar et al., 2012; Dotson et al., 2014). First, broadband signals (sampled at 32kHz) were highpass filtered (Monkey E: 500 Hz – 9 kHz; Monkey L: 500 Hz – 4 kHz). Second, a threshold of 5 standard deviations of the background signal was used to identify spikes. 32 data points were saved for each spike (11 points before and 21 points after and including the minimum). Waveforms were clustered using KlustaKwik (Rossant et al., 2016). Clusters were merged and artifacts were discarded using MClust (http://redishlab.neuroscience.umn.edu/MClust/MClust.html). To be considered a single unit (SUA), waveforms in the cluster were required to be stable over time, non-overlapping with all other clusters, and have an inter-spike interval histogram with a clear refractory period.

Selection of units—Only units with ≥ 1Hz average firing rates were included in the analysis, in order to insure a sufficient amount of activity to perform the firing rate analyses. Also, only areas with a large number of units or ones that could be reasonably pooled with adjacent cortical areas were included. Subsequently, here we report on a lower number of cortical areas than were actually recorded from Dotson et al. (2017). The average firing rate was calculated for each unit using all correct working memory trials, from the presample period to the match onset. Units were considered to have a short latency visual response (SLVR) if they demonstrated a large change in firing rate within 50-100 ms after the sample onset. This indicated that their receptive fields were likely within the region covered by the sample stimulus. This was necessary because recordings were made over large areas of V1 and V2. We analyzed these units separately. See Table S1 for the total number of units and the number of units with a short latency visual response in each area.

QUANTIFICATION AND STATISTICAL ANALYSIS

Firing rate analyses—All data analysis was performed using MATLAB with both custom and built-in code. Only correct trials were analyzed. Results were similar across the two monkeys, so data was pooled. Areas with a low number of sampled units were pooled with adjacent areas to form a total of 24 areas or areal groups (Table S1). Firing rate analyses were performed on each unit, over time, using 200 ms time bins (±100 ms from center of bin), stepped every 50 ms. Time bins went from 250 ms prior to the sample onset to 700 ms after the sample offset (bins 1 to 30), and from 300 ms to 100 ms prior to the match onset (bins 31 to 35). For each analysis, we performed a false discovery rate (FDR) correction over all time bins for each unit individually using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). To confirm that the false discovery rate correction was appropriate, we examined the pre-sample period for each area/ group. This revealed that during both the task-dependent and stimulus-specific analyses, the incidence of significance was nearly always below 5% at all time bins during the pre-sample period (Figures 3, 5, S3, and S4).

The task-dependence of each unit, at each time bin, was determined by comparing the firing rates during the match-to-sample trials to the firing rates during fixation trials, using the Wilcoxon rank sum test ($p < 0.05$). For each area/group, we calculated the incidence at each time bin by dividing the number of significant observations by the number of units. We determined if the differences in firing rates during the dMTS task were enhanced (increased) or suppressed (decreased) with respect to the fixation task by performing a binomial test on the counts of enhanced and suppressed units ($p < 0.05$). Only time bins that had an incidence of task-dependence > 5% were tested. To visualize these results, we calculated the percentage difference of enhanced and suppressed units [((number enhanced − number suppressed) / (number enhanced + number suppressed))*100]. We refer to this measure as the rate modulation index (Figure 2B).

The stimulus-selectivity of each unit, at each time bin, was determined by comparing the firing rates across stimuli, using the Kruskal-Wallis test ($p < 0.05$). For each area/group, we calculated the incidence at each time bin by dividing the number of significant observations by the number of units. To calculate the normalized incidence we simply divided the incidence at each time bin by the sum across all time bins. We then summed the time bins from 500 to 700ms (bins 26-30) after the sample offset and the match locked period (bins 31-35) to determine the relative incidence during the delay period.

To determine the incidence of units that are both task-dependent and stimulus-selective (Figure S5) we simply found the overlap between the two analyses. Specifically, for each unit, at each time bin we determined if the activity was both stimulus-selective and task-dependent.

To identify how long into the task units were recruited, we made a histogram for each cortical area/group of the first time bin that units were stimulus selective. We excluded the presample period (first four time bins) from the analysis. We then calculated the cumulative sum for visualization and to identify the last time bin that new units were recruited. When the cumulative sum reaches one that signifies that all of the units have been recruited (Figure 6C).

To determine if the incidence of stimulus-selectivity in the microsaccade-modulated units was different from the non-microsaccade modulated units, we performed a Pearson's chi-square test of independence at each time bin ($p < 0.05$). Only a small number of the V1 and V2 microsaccade-modulated units were identified as having a short latency visual response (10/98 in V1, and 9/107 in V2). So, we chose to only analyze the V1 and V2 units without short latency responses. For each area, we performed a false discovery rate correction over all time bins using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).

Microsaccade detection procedure—We used an eye movement velocity threshold of 10 dva/s to detect microsaccades. The eye position signals were first lowpass filtered (0-40 Hz) to remove noise. The microsaccades had to last longer than 10 ms and must have occurred at least 50 ms after the previous one (typically they had much longer separations). Since the animals were required to maintain gaze within 1.5 dva of the central fixation dot, the microsaccades were typically < 1dva.

Microsaccade modulation analysis—To identify microsaccade-modulated units, we computed microsaccade locked peri-event time histograms (±200 ms from microsaccade onset, 20 ms non-overlapping bins) using the fixation trials. Since fixation trials used the same timing structure as the match-to-sample trials, we used the "sample offset" time as a reference point for the analysis. Only microsaccades with an onset time > 200 ms and < 600 ms after sample offset were used (Figures S1A and S1B show the main sequences for both animals). To determine if neural activity was significantly modulated by the eye movements, we compared the observed average firing rates to surrogate distributions ($p < 0.05$, two-tailed test). Surrogate distributions were computed by randomizing the trials with respect to the microsaccade times. Figures S1C-S1E show examples of units in areas 8L, V2 and V1 that are modulated by the microsaccadic eye movements. We computed 100 surrogates, and then fit a Poisson function at each time bin in order to estimate p values less than 0.01 (smallest p value using just the surrogate distributions is $1/100 = 0.01$). Each unit was false discovery rate corrected using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). A unit was considered to be microsaccade modulated if > 1 bin was significant. Figure S2A shows the percentage of units in each cortical area/group that are microsaccade-modulated. Figures S2B-S2D show when units in areas 8L, V2, and V1 were modulated with respect to a microsaccade, and if the firing rate was above or below the surrogate distribution. The modulations in area 8L occurred around the time of the microsaccade, consistent with this area's involvement in generating saccadic eye movements. In V1 and V2 there is a suppression followed by an enhancement. These dynamics agree with previous studies (Martinez-Conde et al., 2013) and are likely generated by motor signals to visual cortex rather than retinal input.

Microsaccade pattern analysis—To determine if the microsaccade patterns that occur during the dMTS task are stimulus specific, we used mutual information (MI) analysis of the microsaccade endpoints. MI was calculated using time bins with a 200 ms window (±100 ms from center of bin), stepped every 50 ms. Time bins went from 250 ms prior to the sample onset to 700 ms after the sample offset (bins 1 to 30). At each time bin, we binned the spatial location of microsaccade endpoints into a 10x10 non-overlapping grid that covered 3x3 dva, and then converted these data into a 1D array. This enabled us to easily calculate the mutual information. Mutual information was calculated as follows:

$$MI(S;R) = \sum_{s,r} P(r)P(s \mid r)log_2\frac{P(s \mid r)}{P(s)}$$

where, P(*r*) is the probability of observing the response *r* (microsaccade endpoint), from the response set R (all possible responses); P(*s*) is the probability of stimulus *s* being presented, from the set of stimuli S; P(*s*|*r*) is the posterior probability that the stimulus *s* was presented given the response *r*. For each bin, only one microsaccade was allowed per trial. Since a response did not always occur in every bin and every trial, P(*s*) was adjusted accordingly. This allowed us to determine the information gained about the sample image based on knowing the eye position. We assessed statistical significance by comparing the observed MI values to surrogate distributions (p < 0.05). Surrogates were created by randomizing the trial labels and then computing MI. The same number of trials was kept for each label. Each surrogate distribution (1000 surrogates) was fit with a generalized extreme value function in order to estimate p values less than 0.001 (smallest p value using just the surrogate distribution is 1/1000 = 0.001). Each individual recording session was false discovery rate corrected using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We are grateful to Dr. Chris O'Rourke for her care of the animals and assistance in this study. We thank Tiphani Lynn for her help with the analysis of eye movements and the histological parcellation of cortical areas in monkey E. We thank Elliot Gray for his assistance in the assembly and testing of the microdrives. We thank Susan Krueger for her help in the behavioral training and daily care of the animals. We are grateful to Dr. Rodrigo Salazar and Dr. James Mazer for helpful discussions and comments on an earlier version of the manuscript. We also thank Neuralynx (Bozeman MT) for providing the data acquisition system. This work was supported by grants from the NINDS (R01 NS059312) and the NIMH (R01 MH081162), the McKnight Foundation Memory and Cognitive

Disorders Award, and an EPSCoR RII Track-2 award from the National Science Foundation (1632738) to C.M.G. and S.J.H.

REFERENCES

Andrey P, and Maurin Y (2005). Free-D: an integrated environment for three-dimensional reconstruction from serial sections. J. Neurosci. Methods 145, 233–244. [PubMed: 15922039]

Asaad WF, and Eskandar EN (2008a). Achieving behavioral control with millisecond resolution in a high-level programming environment. J. Neurosci. Methods 173, 235–240. [PubMed: 18606188]

Asaad WF, and Eskandar EN (2008b). A flexible software tool for temporally-precise behavioral control in Matlab. J. Neurosci. Methods 174, 245–258. [PubMed: 18706928]

Baddeley A (2003). Working memory: looking back and looking forward. Nat. Rev. Neurosci 4, 829–839. [PubMed: 14523382]

Benjamini Y, and Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Series B Methodol 57, 289–300.

Bisley JW, Zaksas D, Droll JA, and Pasternak T (2004). Activity of neurons in cortical area MT during a memory for motion task. J. Neurophysiol 91, 286–300. [PubMed: 14523065]

Brandt SA, and Stark LW (1997). Spontaneous eye movements during visual imagery reflect the content of the visual scene. J. Cogn. Neurosci 9, 27–38. [PubMed: 23968178]

Chaudhuri R, Knoblauch K, Gariel MA, Kennedy H, and Wang XJ (2015). A large-scale circuit mechanism for hierarchical dynamical processing in the primate cortex. Neuron 88, 419–431. [PubMed: 26439530]

Christophel TB, Klink PC, Spitzer B, Roelfsema PR, and Haynes JD (2017). The distributed nature of working memory. Trends Cogn. Sci 21, 111–124. [PubMed: 28063661]

Courtney SM, Ungerleider LG, Keil K, and Haxby JV (1997). Transient and sustained activity in a distributed neural system for human working memory. Nature 386, 608–611. [PubMed: 9121584]

D’Esposito M, and Postle BR (2015). The cognitive neuroscience of working memory. Annu. Rev. Psychol 66, 115–142. [PubMed: 25251486]

D’Esposito M, Postle BR, and Rypma B (2000). Prefrontal cortical contributions to working memory: evidence from event-related fMRI studies. Exp. Brain Res 133, 3–11. [PubMed: 10933205]

Dotson NM, Salazar RF, and Gray CM (2014). Frontoparietal correlation dynamics reveal interplay between integration and segregation during visual working memory. J. Neurosci 34, 13600–13613. [PubMed: 25297089]

Dotson NM, Goodell B, Salazar RF, Hoffman SJ, and Gray CM (2015). Methods, caveats and the future of large-scale microelectrode recordings in the non-human primate. Front. Syst. Neurosci 9, 149. [PubMed: 26578906]

Dotson NM, Hoffman SJ, Goodell B, and Gray CM (2017). A large-scale semi-chronic microdrive recording system for non-human primates. Neuron 96, 769–782.e2. [PubMed: 29107523]

Ester EF, Serences JT, and Awh E (2009). Spatially global representations in human primary visual cortex during working memory maintenance. J. Neurosci 29, 15258–15265. [PubMed: 19955378]

Ester EF, Sprague TC, and Serences JT (2015). Parietal and frontal cortex encode stimulus-specific mnemonic representations during visual working memory. Neuron 87, 893–905. [PubMed: 26257053]

Fuster JM, and Jervey JP (1981). Inferotemporal neurons distinguish and retain behaviorally relevant features of visual stimuli. Science 212, 952–955. [PubMed: 7233192]

Fuster JM, Bauer RH, and Jervey JP (1982). Cellular discharge in the dorsolateral prefrontal cortex of the monkey in cognitive tasks. Exp. Neurol 77, 679–694. [PubMed: 7117470]

Harrison SA, and Tong F (2009). Decoding reveals the contents of visual working memory in early visual areas. Nature 458, 632–635. [PubMed: 19225460]

Hasson U, Chen J, and Honey CJ (2015). Hierarchical process memory: memory as an integral component of information processing. Trends Cogn. Sci 19, 304–313. [PubMed: 25980649]

Honey CJ, Thesen T, Donner TH, Silbert LJ, Carlson CE, Devinsky O, Doyle WK, Rubin N, Heeger DJ, and Hasson U (2012). Slow cortical dynamics and the accumulation of information over long timescales. Neuron 76, 423–434. [PubMed: 23083743]

Laeng B, and Teodorescu DS (2002). Eye scanpaths during visual imagery reenact those of perception of the same visual scene. Cogn. Sci 26, 207–231.

Lara AH, and Wallis JD (2015). The role of prefrontal cortex in working memory: a mini review. Front. Syst. Neurosci 9, 173. [PubMed: 26733825]

Lee H, Simpson GV, Logothetis NK, and Rainer G (2005). Phase locking of single neuron activity to theta oscillations during working memory in monkey extrastriate visual cortex. Neuron 45, 147–156. [PubMed: 15629709]

Luck SJ, and Vogel EK (2013). Visual working memory capacity: from psychophysics and neurobiology to individual differences. Trends Cogn. Sci 17, 391–400. [PubMed: 23850263]

Markov NT, Ercsey-Ravasz MM, Ribeiro Gomes AR, Lamy C, Magrou L, Vezoli J, Misery P, Falchier A, Quilodran R, Gariel MA, et al. (2014a). A weighted and directed interareal connectivity matrix for macaque cerebral cortex. Cereb. Cortex 24, 17–36. [PubMed: 23010748]

Markov NT, Vezoli J, Chameau P, Falchier A, Quilodran R, Huissoud C, Lamy C, Misery P, Giroud P, Ullman S, et al. (2014b). Anatomy of hierarchy: feedforward and feedback pathways in macaque visual cortex. J. Comp. Neurol 522, 225–259. [PubMed: 23983048]

Martinez-Conde S, Otero-Millan J, and Macknik SL (2013). The impact of microsaccades on vision: towards a unified theory of saccadic function. Nat. Rev. Neurosci 14, 83–96. [PubMed: 23329159]

Mendoza-Halliday D, Torres S, and Martinez-Trujillo JC (2014). Sharp emergence of feature-selective sustained activity along the dorsal visual pathway. Nat. Neurosci 17, 1255–1262. [PubMed: 25108910]

Miller EK, Li L, and Desimone R (1991). A neural mechanism for working and recognition memory in inferior temporal cortex. Science 254, 1377–1379. [PubMed: 1962197]

Miller EK, Li L, and Desimone R (1993). Activity of neurons in anterior inferior temporal cortex during a short-term memory task. J. Neurosci 13, 1460–1478. [PubMed: 8463829]

Miller EK, Erickson CA, and Desimone R (1996). Neural mechanisms of visual working memory in prefrontal cortex of the macaque. J. Neurosci 16, 5154–5167. [PubMed: 8756444]

Miyashita Y, and Chang HS (1988). Neuronal correlate of pictorial short-term memory in the primate temporal cortex. Nature 331, 68–70. [PubMed: 3340148]

Murray JD, Bernacchia A, Freedman DJ, Romo R, Wallis JD, Cai X, Padoa-Schioppa C, Pasternak T, Seo H, Lee D, and Wang XJ (2014). A hierarchy of intrinsic timescales across primate cortex. Nat. Neurosci 17, 1661–1663. [PubMed: 25383900]

Owen AM, Stern CE, Look RB, Tracey I, Rosen BR, and Petrides M (1998). Functional organization of spatial and nonspatial working memory processing within the human lateral frontal cortex. Proc. Natl. Acad. Sci. USA 95, 7721–7726. [PubMed: 9636217]

Parthasarathy A, Herikstad R, Bong JH, Medina FS, Libedinsky C, and Yen SC (2017). Mixed selectivity morphs population codes in prefrontal cortex. Nat. Neurosci 20, 1770–1779. [PubMed: 29184197]

Peng X, Sereno ME, Silva AK, Lehky SR, and Sereno AB (2008). Shape selectivity in primate frontal eye field. J. Neurophysiol 100, 796–814. [PubMed: 18497359]

Postle BR, Druzgal TJ, and D’Esposito M (2003). Seeking the neural substrates of visual working memory storage. Cortex 39, 927–946. [PubMed: 14584560]

Quintana J, Yajeya J, and Fuster JM (1988). Prefrontal representation of stimulus attributes during delay tasks. I. Unit activity in cross-temporal integration of sensory and sensory-motor information. Brain Res. 474, 211–221. [PubMed: 3208130]

Riggall AC, and Postle BR (2012). The relationship between working memory storage and elevated activity as measured with functional magnetic resonance imaging. J. Neurosci 32, 12990–12998. [PubMed: 22993416]

Rigotti M, Barak O, Warden MR, Wang XJ, Daw ND, Miller EK, and Fusi S (2013). The importance of mixed selectivity in complex cognitive tasks. Nature 497, 585–590. [PubMed: 23685452]

Rossant C, Kadir SN, Goodman DFM, Schulman J, Hunter MLD, Saleem AB, Grosmark A, Belluscio M, Denfield GH, Ecker AS, et al. (2016). Spike sorting for large, dense electrode arrays. Nat. Neurosci 19, 634–641. [PubMed: 26974951]

Salazar RF, Dotson NM, Bressler SL, and Gray CM (2012). Content-specific fronto-parietal synchronization during visual working memory. Science 338, 1097–1100. [PubMed: 23118014]

Sarma A, Masse NY, Wang XJ, and Freedman DJ (2016). Task-specific versus generalized mnemonic representations in parietal and prefrontal cortices. Nat. Neurosci 19, 143–149. [PubMed: 26595652]

Serences JT (2016). Neural mechanisms of information storage in visual short-term memory. Vision Res. 128, 53–67. [PubMed: 27668990]

Serences JT, Ester EF, Vogel EK, and Awh E (2009). Stimulus-specific delay activity in human primary visual cortex. Psychol. Sci 20, 207–214. [PubMed: 19170936]

Sereno AB, and Maunsell JH (1998). Shape selectivity in primate lateral intraparietal cortex. Nature 395, 500–503. [PubMed: 9774105]

Sreenivasan KK, Curtis CE, and D'Esposito M (2014). Revisiting the role of persistent neural activity during working memory. Trends Cogn. Sci 18, 82–89. [PubMed: 24439529]

Supèr H, Spekreijse H, and Lamme VA (2001). A neural correlate of working memory in the monkey primary visual cortex. Science 293, 120–124. [PubMed: 11441187]

Van Essen DC (2012). Cortical cartography and Caret software. Neuroimage 62, 757–764. [PubMed: 22062192]

van Kerkoerle T, Self MW, and Roelfsema PR (2017). Layer-specificity in the effects of attention and working memory on activity in primary visual cortex. Nat. Commun 8, 13804. [PubMed: 28054544]

Zaksas D, and Pasternak T (2006). Directional signals in the prefrontal cortex and in area MT during a working memory for visual motion task. J. Neurosci 26, 11726–11742. [PubMed: 17093094]

Highlights

- Neuronal activity recorded from 42 cortical areas in behaving monkeys
- Task-dependent differences in firing rates are widely distributed
- Stimulus-specific changes in firing rates are hierarchically organized
- Microsaccades during the memory delay encode the stimuli held in memory

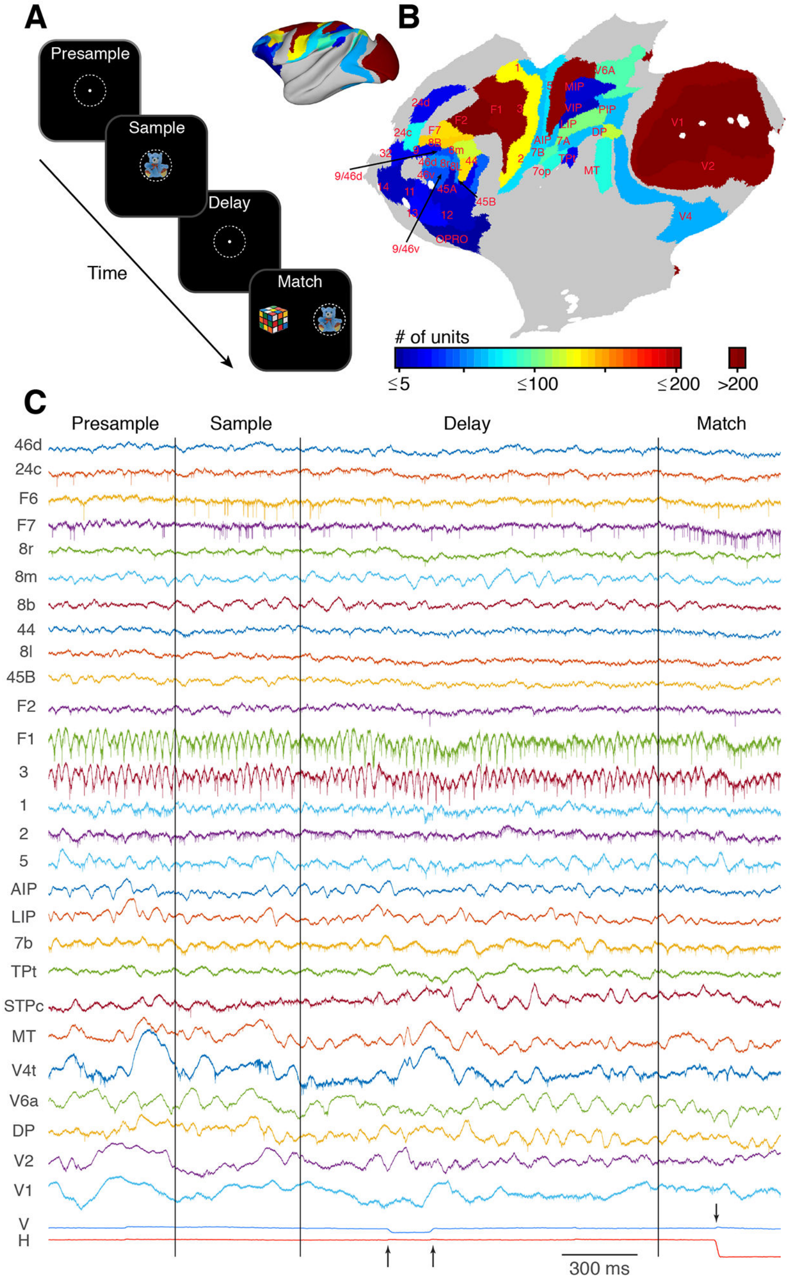

Figure 1. Behavioral Task, Distribution of Recording Sites, and Raw Data Collected on a Single Trial

(A) Schematic of the feature-based delayed match-to-sample task. The monkeys maintained their gaze within a 3° window (dashed white circle) until the match period. A sample stimulus, randomly drawn from a set of 5 possible images, is presented for 500 ms. Following a variable delay, the fixation target is extinguished, and the match stimulus is presented. The match consists of the previous sample image and 1 of the 4 non-sample images presented at 6° eccentricity on either side of the fixation target (STAR Methods). The monkeys are rewarded by making a saccadic eye movement to the sample image.

(B) Cortical flatmap showing the distribution of recorded units in each cortical area. The inset shows the same data on an inflated brain.

(C) Example of the raw broadband data recorded on a single trial of the dMTS task from 27 separate cortical areas in monkey L. The names of each cortical area are shown on the left. The vertical and horizontal eye position signals are shown at the bottom. The vertical lines indicate the times of the sample and the match onset, respectively. The arrows during the delay period indicate two microsaccadic eye movements. The arrow during the match period indicates the time of the choice.

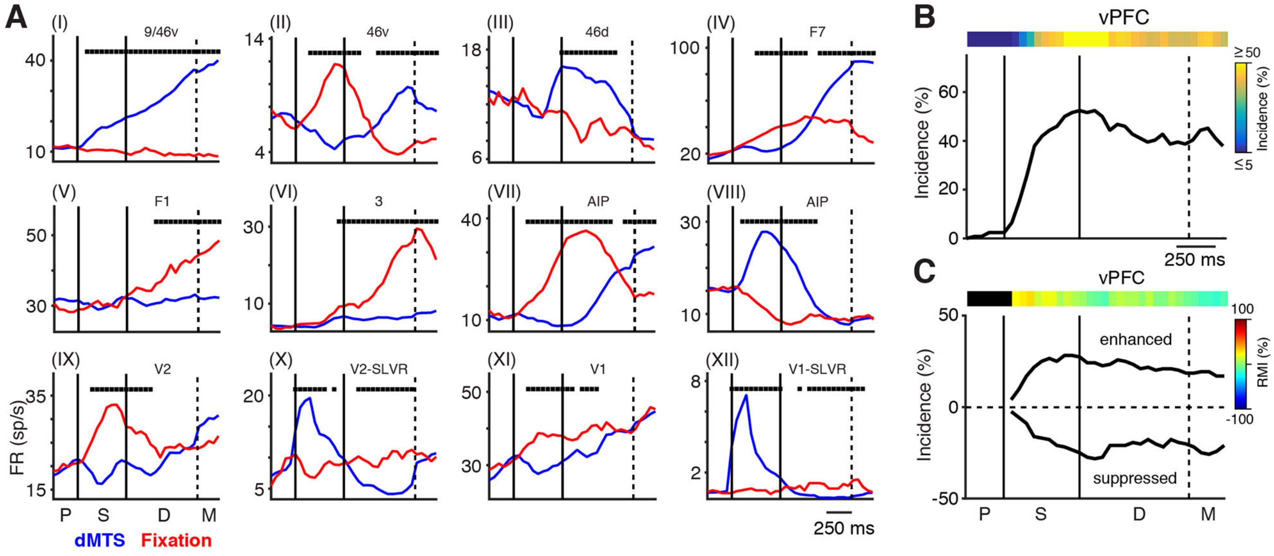

Figure 2. Task-Dependent Changes in Firing Rates

(A) Example PSTHs from units in nine different cortical areas illustrating the average neuronal responses during the dMTS (blue) and interleaved fixation tasks (red) (I–XII). Black squares at the top of each plot indicate the timing of significant differences in firing rates between the two tasks.

(B) Incidence of significant task-dependent activity from all recordings in vPFC in both monkeys as a function of time. The colored bar at the top shows the same data as a heatmap.

(C) The same data as in (B) separated into values in which the responses to the dMTS task are greater (enhanced) or less (suppressed) than those occurring during the fixation task. The colored bar shows the heatmap of the response modulation index. There were no significant differences in the incidence of enhanced and suppressed responses in vPFC. The labels P, S, D, and M indicate the presample, sample, delay, and match locked periods, respectively, and denote the same meaning in Figures 3, 4, 5, 6, 7, and S3-S5.

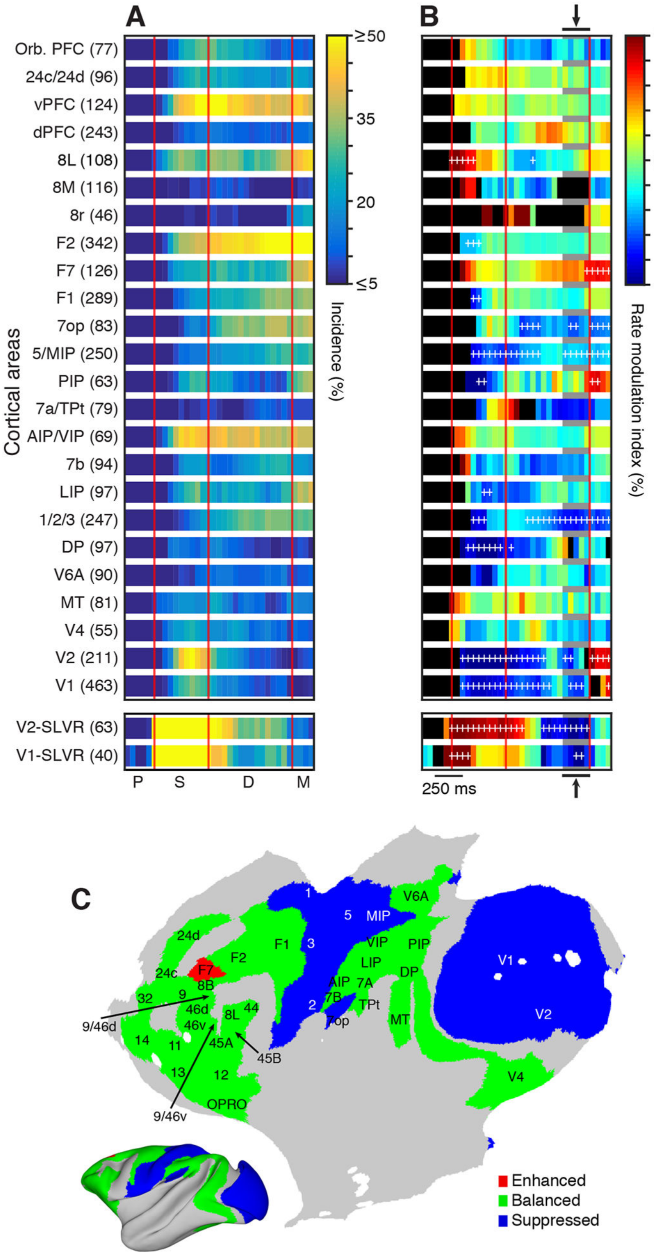

Figure 3. Task-Dependent Unit Activity Is Widely Distributed

(A) Heatmaps showing the time course of the incidence of task-dependent unit activity for each of the 24 areas/groups sampled from both monkeys. The bottom two plots show the data for the units in areas V1 and V2 that displayed short-latency visual responses (SLVR) to at least one of the sample stimuli. Area/group names and number of units recorded are shown on the left.

(B) Heatmaps of the rate modulation index for each area/group over the course of the task. Time bins with an incidence of task dependence <5% are colored black. Positive (warm colors) and negative (cool colors) values indicate that more units have firing ratesduring the dMTS task that are greater (enhanced) or less (suppressed) than the firing rates occurring during the fixation task, respectively. White plus signs mark the bins when the number of enhanced or suppressed units is significantly different.

(C) Flat map showing the general activity pattern during the end of the fixed delay (arrows and gray bars in B).

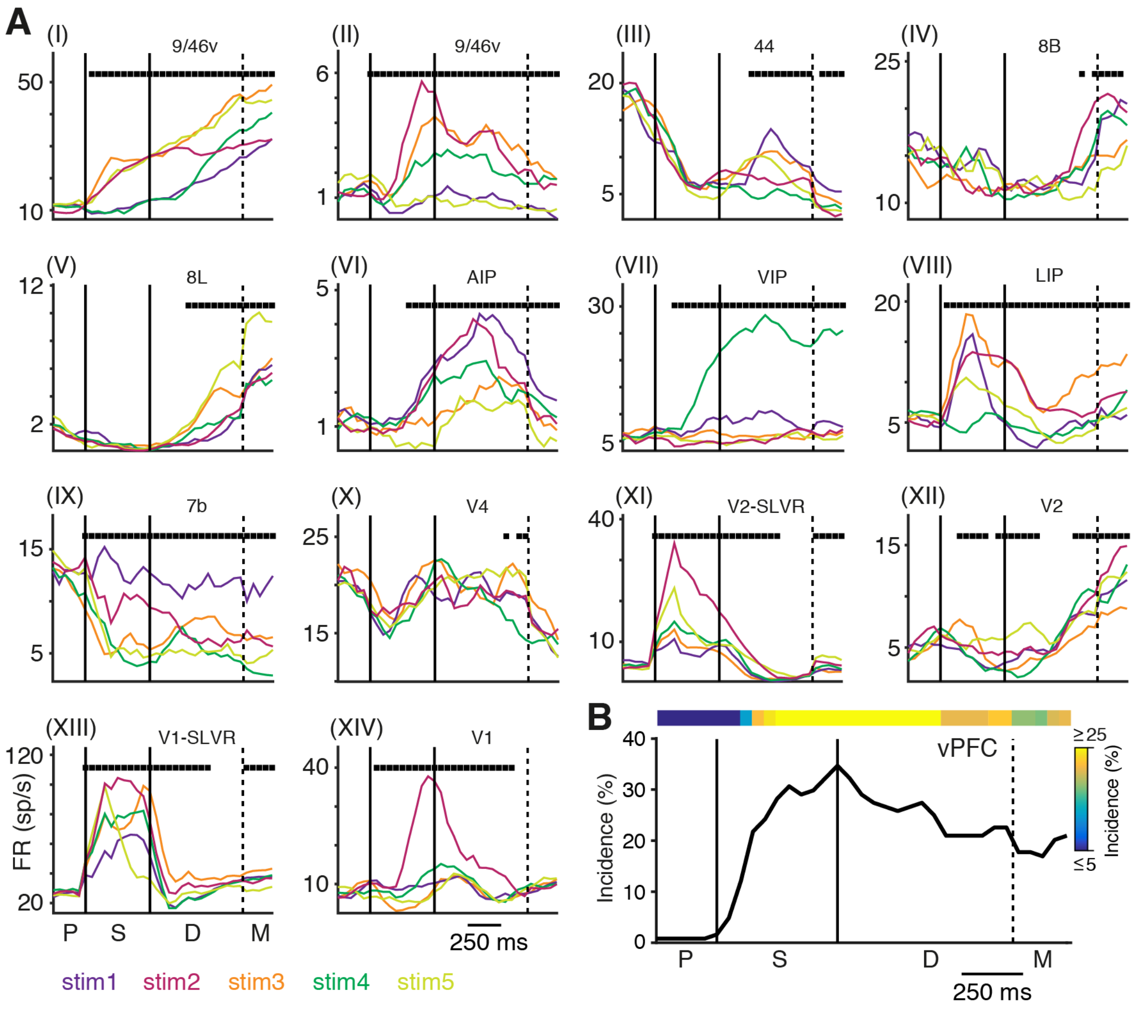

Figure 4. Stimulus-Selective Changes in Firing Rates during the dMTS Task

(A) (I–XIV) Example PSTHs from 11 different cortical areas illustrating the average neuronal responses to each of the 5 sample stimuli during the dMTS task. The response to each stimulus is plotted in a different color (bottom left). Black squares along the top of each plot mark bins where the firing rates are significantly different across stimuli.

(B) Incidence of significant stimulus-selective activity from all recordings in vPFC in both monkeys as a function of time. The colored bar at the top shows the same data plotted as a heatmap.

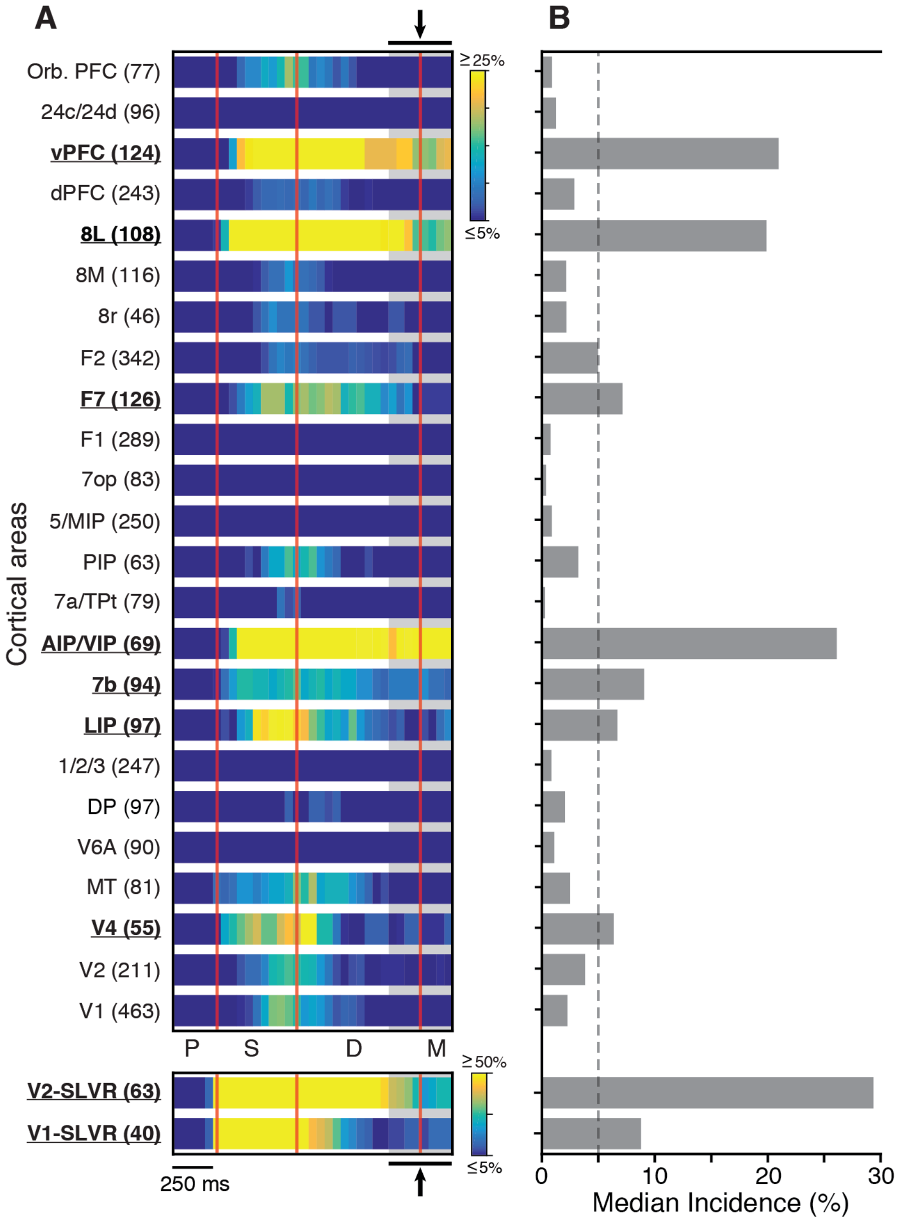

Figure 5. Distribution of Stimulus Selectivity across the Sampled Cortical Areas in Both Monkeys

(A) Heatmaps of the incidence of significant stimulus-selective activity for each area/group over the course of the task. The bottom two plots show the data for the SLVR units in areas V1 and V2. Area/group names and number of units are shown on the left.

(B) Median incidence of stimulus selective activity during the time bins marked by the gray shaded region and arrows (top and bottom) in (A). The dashed line marks the 5% value.

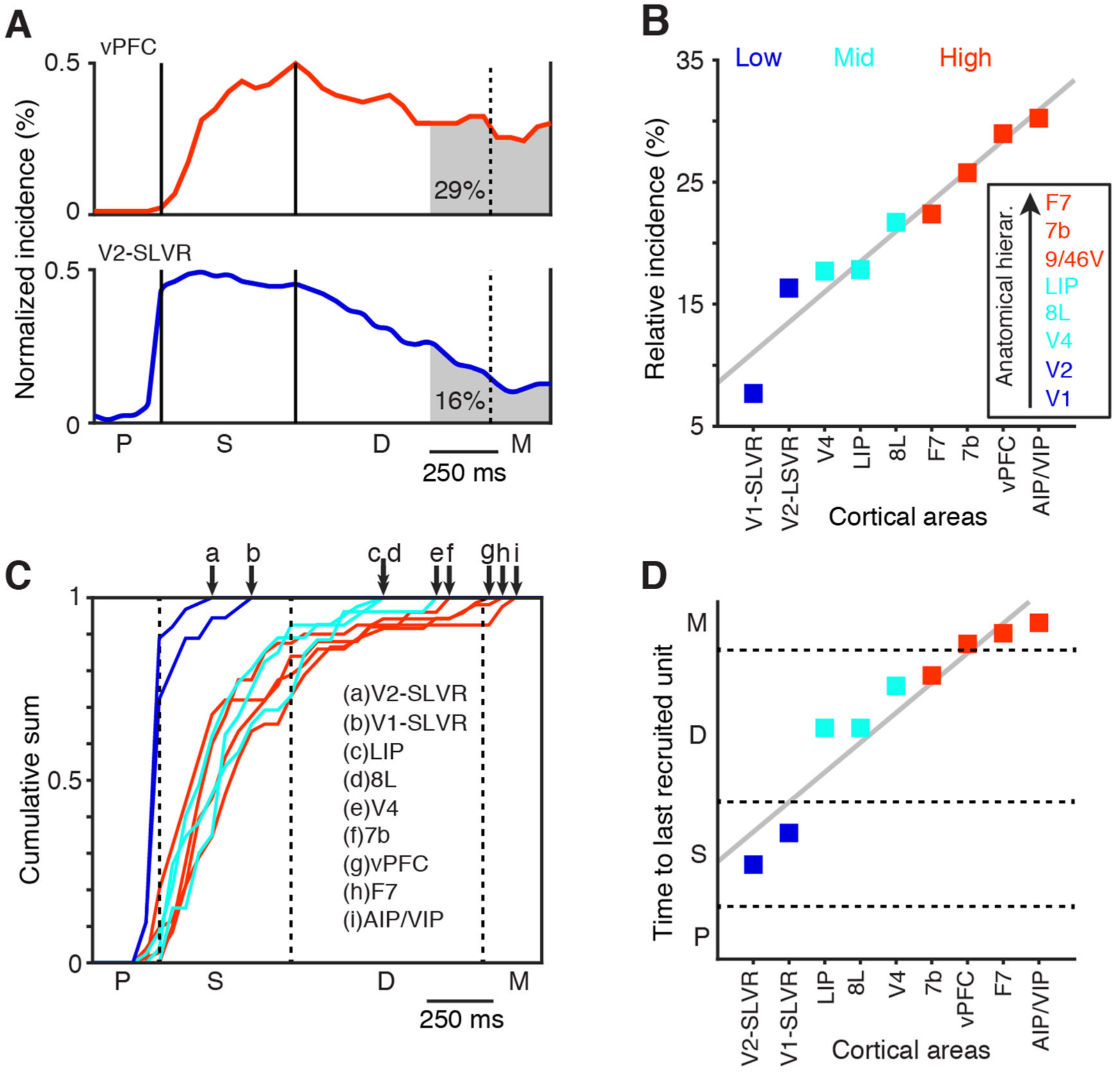

Figure 6. Embedded Hierarchy for Mnemonic Representations

(A) Examples of the incidence of stimulus-specific activity in areas vPFC (top) and V2-SLVR (bottom) (normalized for display purposes). The shaded regions indicate the relative incidence during the late delay period (area under the curve in percentage).

(B) Rank-ordered plot of the relative incidence of stimulus-specific activity during the late delay period, split into three categories: low (blue), mid (cyan), and high (red). The inset shows the anatomical hierarchy derived from Markov et al. (2014b) and Chaudhuri et al. (2015). Areas in the hierarchy are colored based on the same scheme.

(C) Plots of the cumulative sum of the first occurrence of stimulus-specific activity for each of the areas in (B). Arrows at the top indicate the time points where the sum equals 1. The color scheme is the same as in (B).

(D) Rank-ordered plot of the time (vertical axis) when the cumulative sum in (C) reached 1 for each cortical area.

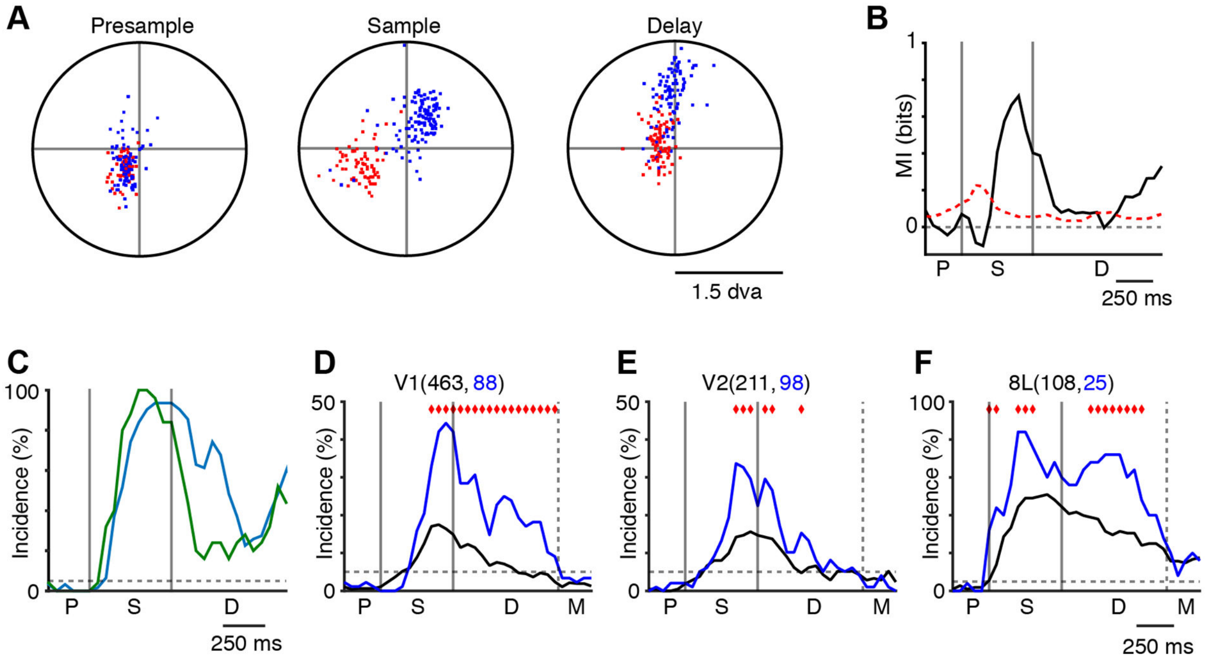

Figure 7. Microsaccades Encode the Sample Stimulus during the Delay Period

(A) Example of microsaccade endpoints (stimulus 1, red dots; stimulus 2, blue dots) during the presample, sample, and delay period for a single recording session from monkey L.

(B) Example of the mutual information (MI) analysis (same data as A, except all five stimuli are used). The MI was bias corrected using the mean of the surrogate distribution. The red dashed line indicates the 95th percentile of the surrogate distribution (also bias corrected).

(C) Summary of the MI analysis, locked to the earliest possible match for both monkeys (monkey E, green; monkey L, blue).

(D–F) Incidence of stimulus-specific activity in V1 (D), V2 (E), and 8L (F) for the MSM units (blue) and the non-MSM units (black) over the course of the task. Red diamonds mark the bins where the incidence values are significantly different. The numbers in parentheses indicate the sample sizes.

Table 1.

List of Areas Contained in Each Group

Group	Areas
Orb. PFC	OPRO, 11, 12, 13, 14, 32
24c/24d	24c, 24d
vPFC	9/46v, 44, 45A, 45B, 46v
dPFC	8B, 9/46d, 9, 46d
8L	8L
8M	8M
8r	8r
F2	F2
F7	F7
F1	F1
7op	7op
5/MIP	5, MIP
PIP	PIP
7a/TPt	7a, TPt
AIP/VIP	AIP, VIP
7b	7b
LIP	LIP
1/2/3	1, 2, 3
DP	DP
V6A	V6A
MT	MT
V4	V4
V2	V2
V1	V1

The first column indicates the group name, and the second column indicates the areas that compose each group. Many of the groups are simply one area. Area names follow Markov et al. (2014a).

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
MATLAB	MathWorks	https://www.mathworks.com/
Solidworks	Dassault Systems	https://www.solidworks.com/
Caret	Van Essen, 2012	http://brainvis.wustl.edu/wiki/index.php/Main_Page
MonkeyLogic	Asaad and Eskandar (2008a, 2008b)	http://www.brown.edu/Research/monkeylogic/
Cheetah	Neuralynx	https://neuralynx.com/
Other		
256-Channel Digital Lynx System	Neuralynx	https://neuralynx.com/
Microelectrodes (Tungsten-in-Glass): exposed end (0.1”), wire dia (.005,” 125 μm), glass shaft dia. (.0098,” 250 μm), 60° taper angle, impedance at 1kHz (~ 1.0 MΩ)	Alpha Omega	https://www.alphaomega-eng.com/