Differential brain mechanisms of selection and maintenance of information during working memory

Romain Quentin[1,3], Jean-Rémi King[2], Etienne Sallard[1], Nathan Fishman[1], Ryan Thompson[1], Ethan Buch[1], Leonardo G Cohen[1]

1. NINDS, Bethesda, USA; 2. New York University, Frankfurt Institute for Advanced Studies, Germany; 3. Lead Contact

Abstract

Working memory is our ability to temporarily hold information as needed for complex cognitive operations. Models of working memory distinguish two separate processes: (i) a selection rule that identifies the content to be recalled and (ii) the maintenance of the content. We aimed to characterize the spatiotemporal neural dynamics underlying these two components. Healthy participants performed a visual working memory task during magnetoencephalography (MEG) recording. Multivariate Pattern Analysis (MVPA) and source analyses identified two distinct types of working memory neural processes underlying selection and maintenance of the content. The selection rule is specifically decoded from sustained low-frequency (<20Hz) neural activity within a cortical network that includes the ventrolateral prefrontal cortex. By contrast, working memory content is transiently reactivated over a distributed and occipito-temporal network that differs from that encoding the sensory stimulus. These results reveal different neural mechanisms that select and maintain information in memory and could account for previous paradoxical reports of persistent and dynamic neural correlates of working memory.

Introduction

Working memory enables the brief holding of information (Baddeley, 2010; Baddeley and Hitch, 1974) and is crucial for a wide range of cognitive tasks in everyday life (Klingberg, 2010). For example, while driving a car, previous visual input providing important contextual information must be maintained for several seconds in order to act appropriately. The same mechanism applies when conversing with a friend, watching a movie or learning a motor skill. Despite the central role of working memory in complex behaviors, how the brain selects and maintains memory content remains actively debated (Christophel et al., 2017).

Lesion studies have pointed to the prefrontal cortex as a crucial brain region mediating working memory (Bauer and Fuster, 1976; Jacobsen, 1935; Petrides, 2005). For example, it has been proposed that working memory engages a fronto-parietal neural network similar to that identified during selective attention (Pollmann and von Cramon, 2000). Consistently, results from intracranial recordings in monkeys and neuroimaging studies in humans have shown that persistent neural activity within prefrontal regions supports working memory (Courtney et al., 1998; Funahashi et al., 1989; Fuster and Alexander, 1971; Goldman-Rakic, 1995).

On the other hand, maintenance of information during working memory appears to engage different brain regions depending on the type of information (Christophel et al., 2012; Ester et al., 2015; Han et al., 2013; Harrison and Tong, 2009; Lee and Baker, 2016). For example, maintenance of visual orientation information engages early visual areas (Riggall and Postle, 2012), while the maintenance of single auditory tones engages the auditory cortex (Kumar et al., 2016). Importantly, such maintenance may not require persistent neural activity (Stokes, 2015). Electrophysiological findings showed that this working memory process depends on highly specific neural temporal dynamics (Fuentemilla et al., 2010; Lundqvist et al., 2016; Stokes et al., 2013).

While early work proposed that neurons in the lateral prefrontal cortex store working memory information (Funahashi and Kubota, 1994; Fuster and Alexander, 1971), recent evidence suggested that prefrontal cortex activity exerts top-down influences on sensory regions, reflecting the selection of information for goal-directed behavior (Curtis and D'Esposito, 2003; Riggall and Postle, 2012; Warden and Miller, 2010). In this framework, working memory depends on two processes: (i) a selection rule that identifies the relevant content to be recalled and (ii) the maintenance of this content for future processing (Vogel et al., 2005). Accordingly,

during a working memory task, a retrospective cue instructing the participant to select one specific memory content increases working memory performance for that item (Griffin and Nobre, 2003; Murray et al., 2013). Yet, most studies do not dissociate the neural responses to the selection from those of the maintenance component and only investigate the spatial localization of working memory content (Christophel et al., 2017).

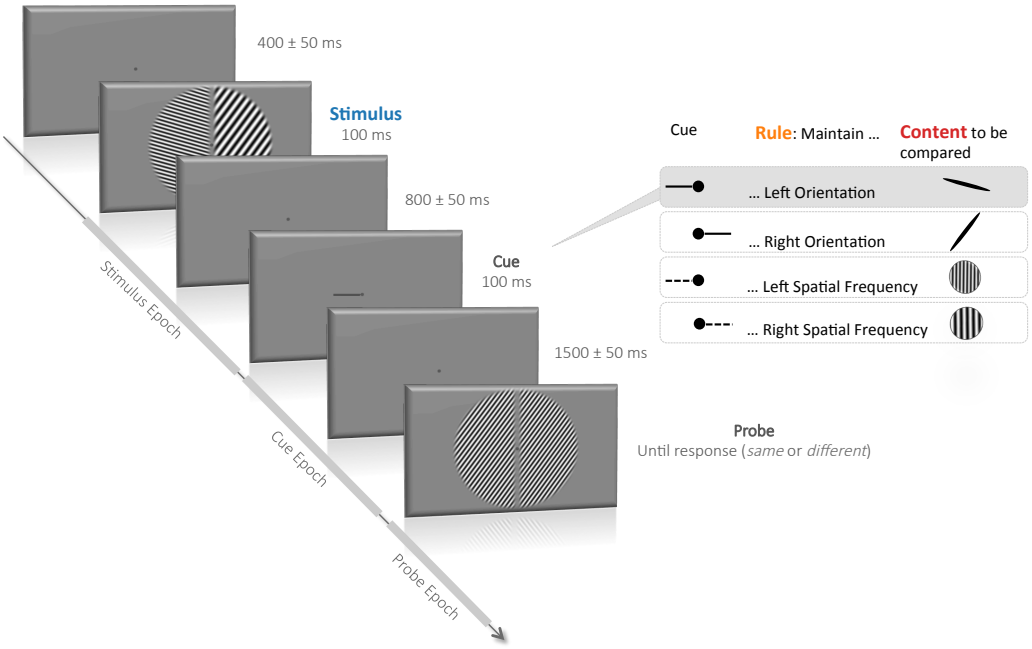

Figure 1: Visual working memory task. A fixation dot appears in the center of the screen and the participant is instructed to fixate this dot. After 400 ± 50ms, the stimulus appears for 100 ms and is composed of four possible different visual attributes: left and right spatial frequency (each chosen among five possible: 1, 1.5, 2.25, 3.375 or 5.06 cycles/degree) and left and right orientation (each chosen among five possible: -72, -36, 0, 36, 73 in degree, 0 being the vertical). After a delay of 800 ± 50ms, the cue appears for 100ms and indicates which visual attribute of the stimulus the participant has to compare with the upcoming probe. A left or right solid line cue indicates respectively the left or right orientation and a left or right dotted line indicates respectively the left or right spatial frequency of the stimulus. After a 1500 ± 50ms delay, the probe appears and the participant is required to answer with one of two fingers whether the cued stimulus attribute is the same or different than the corresponding probe attribute. The probe displays only one orientation and one spatial frequency. In the trial depicted in the figure, the solid line cue pointing to the left instructs the participant to compare the orientation on the left side of the stimulus with the orientation in the probe (the correct answer in this case is "different"). We refer to the time between the stimulus and the cue as the stimulus epoch, the time between the cue and the probe as the cue epoch and the time after the probe as the probe epoch.

Here, we investigated the neural mechanisms that allow our brain to select, transform and maintain a sensory stimulus during a working memory task. We addressed this question by designing a working memory task that enabled the dissociation of the neural mechanisms underlying selection rule from those underlying maintenance of the memory content.

Multivariate pattern analysis (MVPA) of time-resolved brain MEG activity identified two distinct neural mechanisms underlying these two working memory components: the selection process is encoded in a stable low-frequency neural activity within a network that includes the ventrolateral prefrontal cortex while the memory content is transiently reactivated over a distributed and posterior network different from that encoding the sensory stimulus.

Results

We recorded MEG in 23 participants while they performed a working memory task. Each trial started with the visual presentation of a four-dimensional stimulus with two distinct visual gratings (left and right) that varied in line orientation and spatial frequency. A small retrospective visual cue presented ~900ms after the stimulus onset indicated the visual attribute to be retained for a subsequent probe. Specifically, a small line indicated the side (left or right) and the feature (orientation or spatial frequency) of the stimulus to be remembered. Finally, participants indicated whether the cued attribute matched the corresponding attribute of a visual probe presented ~1500ms after the cue onset (**Fig 1**). To isolate the neural representation encoding the selection rule and the memory content, we applied MVPA to decode the four visual attributes of the stimulus (orientation and spatial frequency of each visual grating), the selection rules (spatial and feature rule) and the memory content during the stimulus epoch (-0.2 s to 0.9 s around stimulus onset), the cue epoch (-0.2 s to 1.5 s around the cue onset) and the probe epoch (-0.2 s to 0.4 s around the probe onset; **Fig 2**).

Parallel and transient encoding of four visual attributes

Left and right spatial frequencies could be decoded from 33 ms and 25 ms after stimulus onset respectively (cluster level, $p<0.05$ corrected). The decoding performance peaked around 50ms and rapidly decreased afterwards but remained above chance throughout most of the stimulus epoch. Mean spatial frequency decoding performance over the stimulus epoch was significantly above chance (both $p<0.001$). By contrast, these visual attributes could not be decoded during the cue or the probe epochs. Similar results were observed for the decoding of the left and right orientation. Specifically, orientation decoding started approximately 46 ms after stimulus onset, peaked around 100 ms and remained above chance throughout most of the stimulus epoch. Mean orientation decoding performance was significantly above chance during the stimulus epoch (both $p < 0.001$). Very weak but still significant decoding was also observed during the cue (right orientation: $p < 001$, left orientation: $p < 0.01$) and the probe epochs (both $p < 0.01$)

(**Fig 2A**). Similar decoding results were observed in the time-frequency domain with significant decoding clusters during the first 400 ms after the stimulus onset (**Fig 2B**).

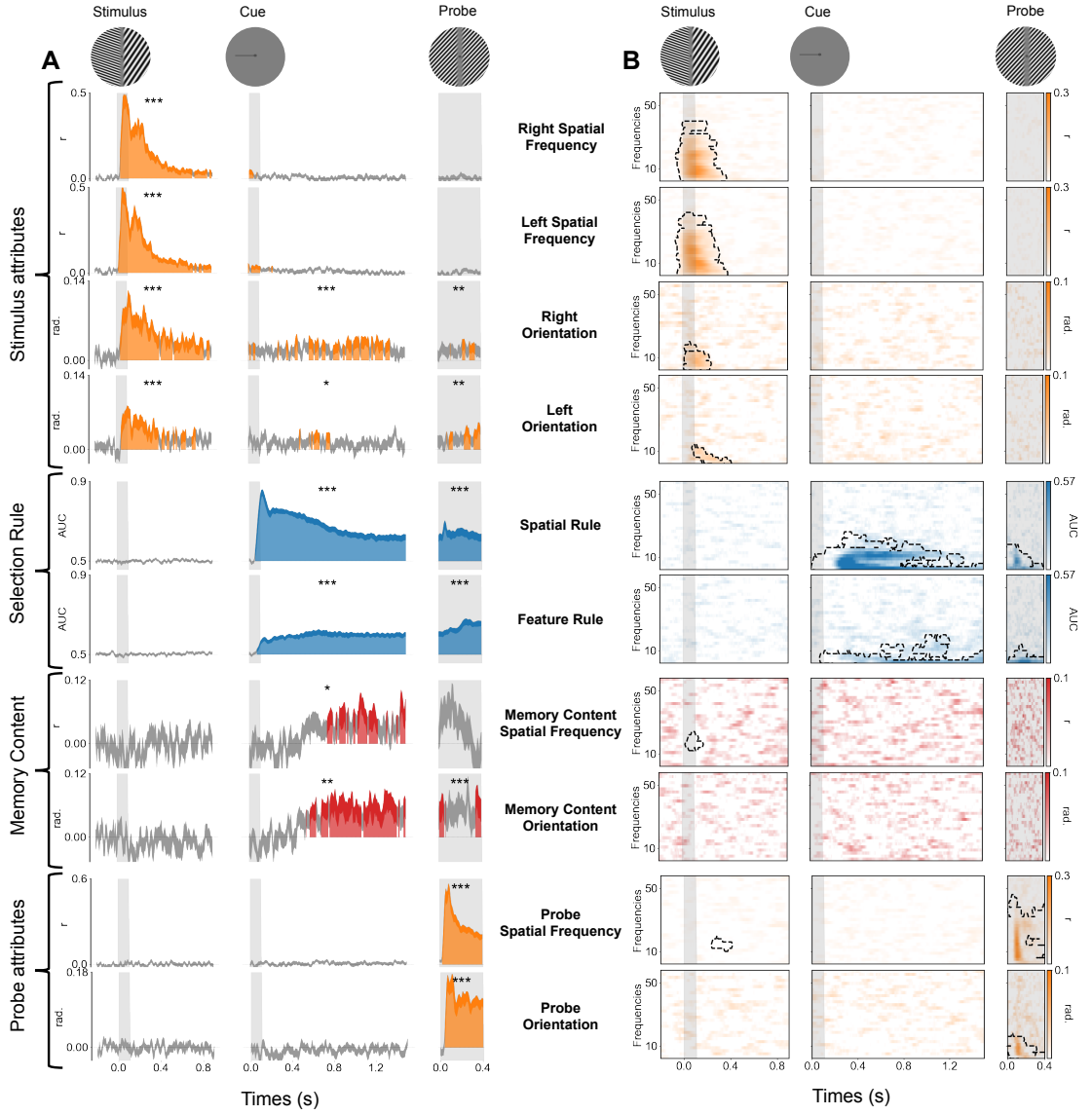

Figure 2: Neural dynamics of visual perception, selection rule and memory content in evoked and time-frequency domains. A. Time course of MEG decoding performance. The x-axis corresponds to the time relative to each event (stimulus, cue and probe, see top) and the y-axis corresponds to the decoding performance for the stimulus attributes, the selection rule, the memory content and the probe attributes. Vertical gray bars indicate the visual presentation of each image (stimulus, cue and probe). Color filled areas depict significant temporal clusters of decoding performance (cluster-level, $p<0.05$ corrected). Variance (thickness of the line) is shown as standard error of the mean (SEM) across participants. Note the successful decoding of the four visual attributes of the stimulus, the spatial and feature rule, the memory content (cued - uncued) for both spatial frequency and orientation and for the two attributes of the probe. The asterisks indicate the significance of the mean decoding performance over the entire corresponding epoch (***, **, * indicate respectively $p < 0.001$, $p < 0.01$ and $p < 0.05$). B. Decoding performance in the time-frequency domain. The x-axis corresponds to the time relative to each event (stimulus, cue and probe, see top) and the y-axis depicts the frequency of MEG activity (between 2-60Hz). Significant clusters of decoding performance are contoured with a dotted line. Note

the successful decoding in the time-frequency domain of the four visual attributes of the stimulus, both the spatial and the feature rule and the two attributes of the probe but not the memory content.

To estimate the brain sources underlying these decoders, the MEG signal was reconstructed in the source space at a single trial level and the same decoding analyses were performed on the source signal. The weights of the estimators were transformed into interpretable patterns of activity (Haufe et al., 2014). The source pattern of activity indicated that the calcarine, the cuneus and lateral occipital regions encoded this information (**Fig 3A**). Overall, our decoding results during visual perception confirmed that multiple visual attributes are simultaneously encoded in the early neural response for several hundred milliseconds, but rapidly become undetectable after about one second (**Fig 2A**).

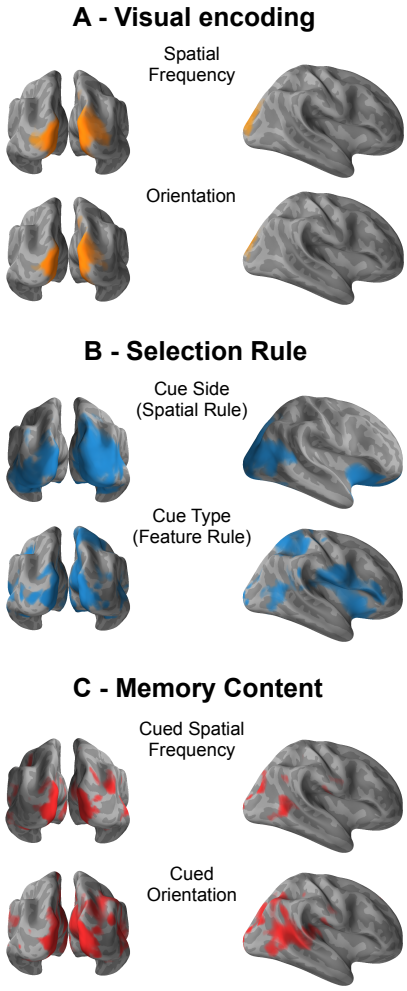

Figure 3: Spatial source representation of stimuli, selection rule and memory content. **A.** Encoding of visual attributes during the stimulus epoch. The calcarine cortex, the cuneus and lateral occipital regions encoded the visual attributes of the stimulus during the stimulus epoch. **B.** Selection rule during the cue epoch. A large cortical network including the ventrolateral prefrontal regions encoded the selection rule. **C**. Memory content during the cue epoch. The neural representation of memory content involves an occipitotemporal brain network.

The selection rule is encoded in stable oscillatory activity involving the ventrolateral prefrontal cortex.

The cue side (spatial rule) and the cue type (feature rule) could be decoded shortly after the cue presentation and during the entire cue and probe epochs (**Fig 2A** and **Fig 4**). The cue side and cue type were decoded respectively 58ms and 75ms after cue onset (cluster level, $p<0.05$ corrected) and the decoding performance remained above chance throughout the cue and probe epochs (**Fig 2A**). To ensure that these decoded patterns of brain activity corresponded to the selection rule and not to the sensory features of the cue, we decoded the same visual cue in a one-back control task. In the initial 200 ms following cue onset, decoding performance of cue side and type were comparable in both tasks (with and without associated selection rule). Subsequently, decoding was significantly higher in the working memory condition than in the control one-back task (**Fig4**, left panel). Overall these decoding results demonstrate that the sustained activity encoding the cue is specific to the selection rule.

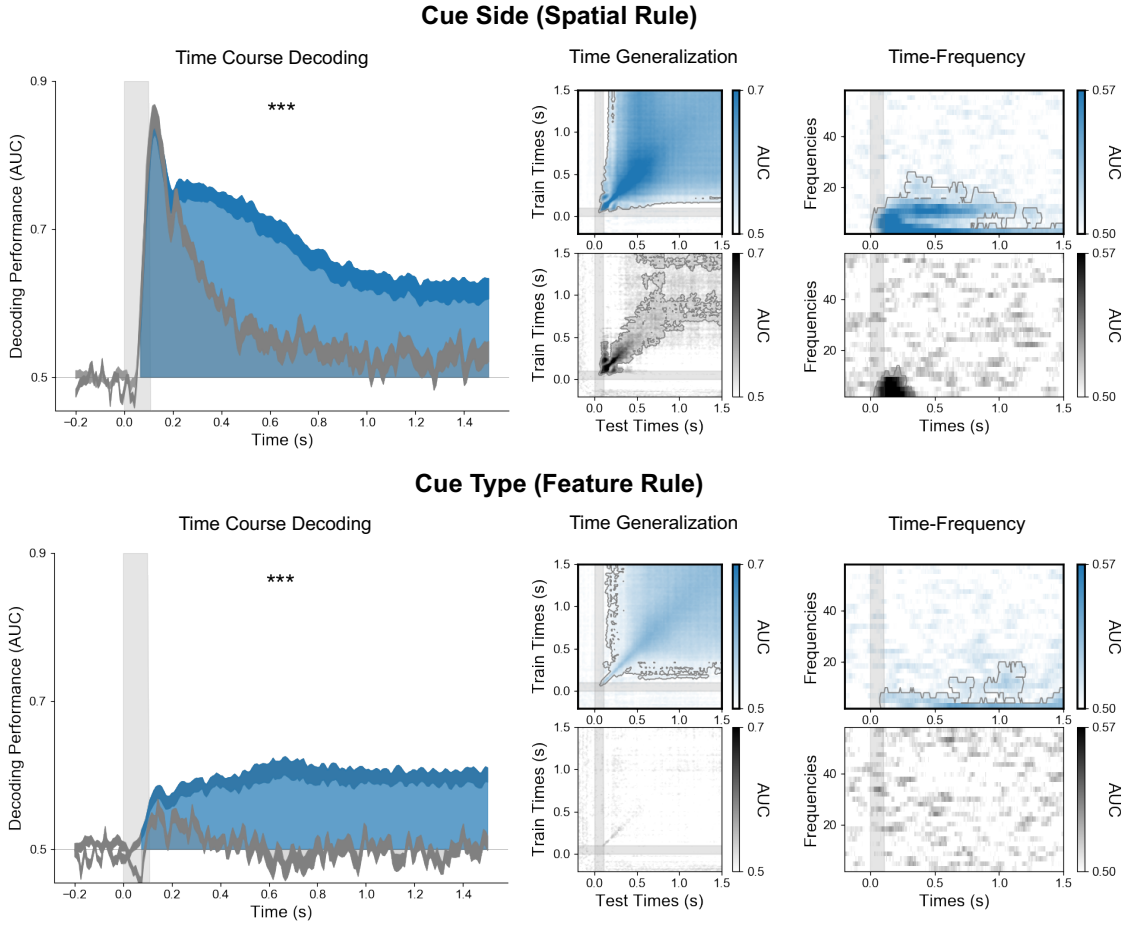

Figure 4: The selection rule is encoded in a persistent and stable pattern of low-frequency brain activity. Neural temporal dynamics of rule selection in evoked and time-frequency domains. On the left, time course (x-axis) of decoding performance (y-axis) during the cue epoch for the cue side (top)

and the cue type (bottom) during the working memory task (blue) when the cue is associated with the selection rule and the control one-back task (gray) when it is not. Note that decoding performance was significantly higher in the working memory task than in the control one-back task. The time generalization matrices (middle panels), in which each estimator trained on time *t* was tested on its ability to predict the variable at time *t'*, identified stable neural representations for both spatial and feature rules. The right panel shows the decoding in the time frequency domain. Note that both rules are maintained within low frequency bands alpha (~10Hz) and delta (~3Hz) activity.

To test the dynamics of the neural representation of the rule, each estimator trained on time *t* was tested on its ability to predict the variable of interest at time *t'* (King and Dehaene, 2014). This temporal generalization analysis showed very stable neural representations for both spatial and feature rules (**Fig 4**, middle panel).

To investigate the oscillatory component of the selection rule, we computed the time-frequency decomposition at the single trial level and applied MVPA on the power in each frequency band. The cue side and the cue type during the working memory task could be decoded in the alpha (~10hz) and delta (~3Hz) bands peaking after the cue onset and until the end of the cue epoch ($p<0.05$ corrected). In the control one-back task, the cue side could be decoded in the frequency domain only for a short period after the cue onset and the cue type could not be decoded at all (**Fig 4**, right panel).

Finally, the same decoding analyses were also performed on the source signal (**Supp Fig 2**). Both spatial and feature rules were encoded in a large network involving the ventral prefrontal, parietal and occipital cortices (**Fig 3B**). Specifically, the activity pattern for the spatial rule involved bilateral orbitofrontal regions, bilateral insula, bilateral inferior parietal lobules, right superior parietal and temporo-parietal junction and bilateral occipital regions and fusiform areas. The activity pattern for the feature rule involved the right orbitofrontal region, inferior frontal gyrus and insula, bilateral peri-central regions, the right superior parietal lobule, bilateral middle temporal regions and bilateral occipital regions including the fusiform area. Overall, our decoding and source results showed that the selection rules (both spatial and feature rule) are associated with sustained oscillatory neural activity and involve ventrolateral prefrontal cortex.

<u>The memory content is transformed and encoded in transient neural activity in a distributed posterior network.</u>

Decoding performance for the memory content (cued orientation and cued spatial frequency) started around 500 ms after the cue onset ($p<0.05$ corrected) and remained above chance

throughout the cue epoch. The mean decoding performance was significantly above chance during the cue epoch both for the cued orientation ($p<0.001$) and cued spatial frequency ($p<0.01$; **Fig 5A**). The working memory content could not be decoded in the time frequency domain (**Fig2B**). Temporal generalization analyses showed a partially stable representation over time for both items (**Fig 5A**). By contrast, the un-cued orientation or the un-cued spatial frequency could not be decoded (**Fig 5B**).

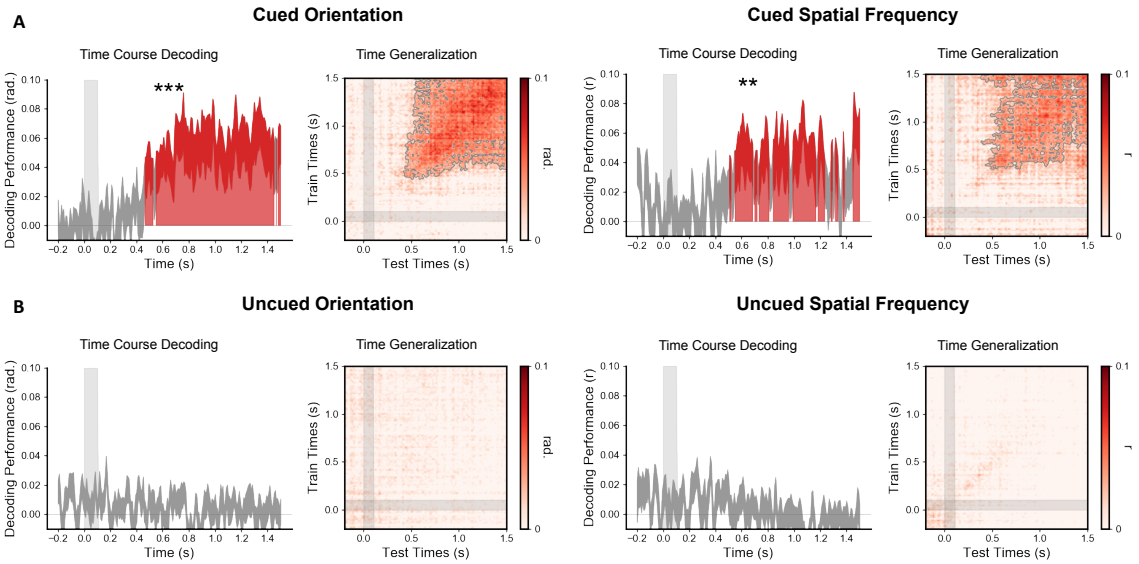

Figure 5: The memory content is transiently reactivated 500 ms after the cue.
A. Time course of decoding performance (y-axis) during the cue epoch for the cued orientation (5 possible orientations) and the cued spatial frequency (5 possible spatial frequencies) during the working memory task and their corresponding time generalization analysis. **B**. Same analysis for the uncued orientation and spatial frequency. Note that decoding performance was significant for the cued but not uncued orientation and spatial frequency. Additionally, decoding was significantly higher for the cued than the uncued item (see **Fig 2** for this difference). In the time-frequency domain, MVPA did not decode any significant clusters for the memory content (not shown here, see **Fig 2**).

The source pattern of activity representing memory content (either orientation or spatial frequency) showed a distributed and posterior network involving bilateral occipital regions, bilateral inferior temporal and temporo-parietal junctions, bilateral posterior temporal regions and left premotor areas (**Fig 3C**). To test if the neural representation of the memory content was similar to that of visual encoding, we tested the generalization across condition. Specifically, we trained the decoders on the visual attributes during perception of the stimulus and tested their ability to decode the memory content, and, conversely, we trained the estimators on the memory content and tested their ability to decode visual attributes during visual encoding. These cross-condition decoding analyses revealed no above chance decoding (**Fig 6** and **Supp**

Fig 5) demonstrating that the neural representation of the memory content differs from the representation of the same attribute during sensory encoding (for full temporal generalization of these cross-condition decoding, see **Supp Fig 5**).

Discussion

We used time-resolved MVPA analysis of MEG data to investigate how the brain selects and maintains information in working memory. The time-course, time-frequency and source-space features emerging from our decoding analyses showed different spatiotemporal neural dynamics for the selection and for the maintenance of information in working memory. In each trial, the task involved visual encoding of four visual attributes, a retrospective cue indicating one of these visual attributes and finally a probe to match with the cued item. We report that the visual attributes of the stimulus were simultaneously encoded in visual brain regions over a period lasting approximately one second after the stimulus onset and then became undetectable or barely detectable. Thus, we showed that MEG signal was rich and spatially selective enough to simultaneously decode the four different visual attributes (**Fig 2**), extending previous findings (Cichy et al., 2015; Stokes et al., 2015). The selection rule was encoded in sustained frequency-specific neural activity in a network that includes the lateral prefrontal cortex. Then, the representation of the memory content was transformed into a transient activity pattern, qualitatively different from its neural representation during initial visual encoding, within a distributed posterior network.

Working memory selection

The experimental paradigm allowed us to identify the representation of two different selection rules, a spatial one indicated by the cue side and a feature one indicated by the cue type. Both rules share similar spatiotemporal neural properties: a very stable neural representation demonstrated by the time generalization (**Fig 4**), a low-frequency oscillatory mechanism demonstrated by the time-frequency decoding (**Fig 2 and 4**) and the involvement of the ventrolateral prefrontal and occipito-parietal regions (**Fig 3**). Our time generalization results are consistent with those obtained using MVPA on intracranial recordings in primates during the period where monkeys needed to maintain a rule (Stokes et al., 2013). In a biological system, a stable representation over time is likely to be more easily readable by a third party (another brain region) than a constantly changing representation that would require continuous shifting of readout algorithms (Murray et al., 2017). The brain oscillatory signature of the spatial and feature rules identified here are reminiscent of the neural signatures of spatial attention, which

engage alpha (Sauseng et al., 2005; Worden et al., 2000) and beta or low-gamma (Buschman and Miller, 2007; Phillips and Takeda, 2009) brain oscillatory activity. It is also in line with reported oscillatory synchronization of local field potentials representing the rule in monkeys (Buschman et al., 2012). The involvement of the ventrolateral frontal cortex is not surprising given its recognized role mediating top-down influences (Sreenivasan et al., 2014) and its contribution to rule representation (Reverberi et al., 2012; Woolgar et al., 2011) and active selection (Petrides, 1996).

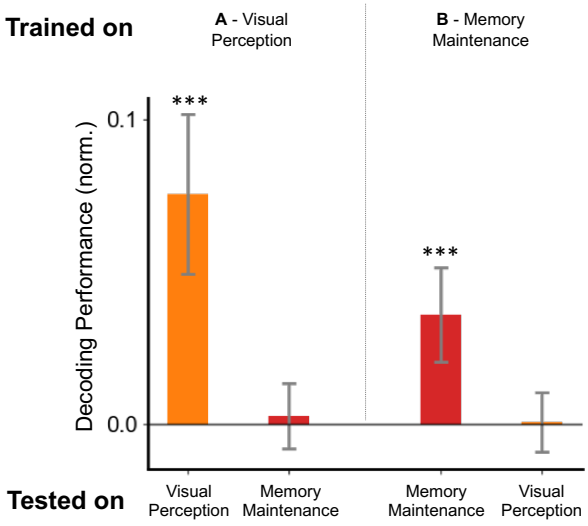

Figure 6: Different neural representations of memory and perceptual content. **A**. Average decoding performance when estimators are trained on the stimulus attributes (orientation or spatial frequency) during the stimulus epoch and either tested during the same epoch or tested on the corresponding memory content during the cue epochs. Note that an estimator trained to decode a visual feature during perception cannot decode the corresponding memory content during the cue epoch. **B**. Average decoding performance when estimators are trained on the memory content (orientation or spatial frequency) during the cue epoch and either tested during the same epoch or tested on the corresponding stimulus attribute during the stimulus epochs. Note that an estimator trained to decode a memory content during the cue epoch cannot decode the corresponding stimulus feature during perception (see **Supp Fig 5** for full time generalization).

Working memory maintenance

Subtle visual differences in the spatial frequency and line orientation of the memory content were decoded a few hundred milliseconds after the cue onset (**Fig 4**). The time generalization suggests that the memory content had a more stable neural representation than the initial visual encoding, even if this result must be interpreted cautiously as a jitter across trials in the reactivation of the memory content could lead to a generalization pattern similar to the one we observe here (Vidaurre et al., 2018). Functional MRI studies using MVPA have shown that the memory content can be decoded from a wide range of brain regions, including occipital (Harrison and Tong, 2009; Serences et al., 2009), parietal (Christophel et al., 2012), temporal (Han et al., 2013) and frontal (Ester et al., 2015) areas. These findings suggested that the brain

regions maintaining the memory content depend on its specific feature, *e.g.*, orientation in early visual areas, motion in extrastriate cortex including area MT+ , or single tones in auditory cortex (Emrich et al., 2013; Kumar et al., 2016; Riggall and Postle, 2012). It has also been shown that the level of abstractness influence the spatial localization of the memory content, with low-level sensory features being encoded in sensory areas and more abstract representations in anterior frontal regions (Christophel et al., 2017; Lee et al., 2013). Thus, it has been proposed that the system maintaining the information in working memory may be the same as the one involved in the encoding of this information (D'Esposito and Postle, 2015). Our source results indicate that the memory content is maintained in a distributed network involving posterior brain regions that include sensory visual areas (**Fig 3**). However, our results also show that the neural representations of the same content during perception and memory delay are qualitatively different, as demonstrated by the lack of generalization across conditions (**Fig 6**). Differentiated representations of memory and visual perception is likely to result in more stable and resistant-to-interference memory content than if they were sharing the same neural substrate (Makovski et al., 2008). Altogether, these findings are consistent with the view that, once the brain knows the rule to apply, it reformats the cued visual representation for a later use at the time of the probe (Myers et al., 2017).

Decoding of the stimulus visual attributes fell back to chance level about 900 ms after the stimulus onset and reappeared following a silent neural activity period about 500ms after the cue. It is possible that short-term changes in synaptic weights in the absence of persistent neural activity are enough to underlie the maintenance of information in working memory (Lewis-Peacock et al., 2012; Stokes et al., 2013), resulting in "activity-silent" neural states (Stokes, 2015). Interestingly, such states could theoretically be reactivated by probing the brain with a light flash (Wolff et al., 2017) or TMS pulse (Rose et al., 2016), as a result of a matched filter mechanism (Sugase-Miyamoto et al., 2008). However, it should be kept in mind that the absence of MVPA decoding does not prove the absence of content-specific neural activity, and thus we cannot rule out the possibility that this activity silent period simply reflects the MVPA limitation to decode low-level neural activity. Also, we noted a boost in the decoding performance of the spatial and feature rules after the probe onset (**Fig 2**), demonstrating that the rule is reactivated by the new visual input.

Persistent and dynamic nature of working memory

Neurophysiological experiments in primates have identified neurons that remain active during the memory delay in prefrontal (Funahashi et al., 1989; Fuster and Alexander, 1971) and other cortical and subcortical regions (Chelazzi et al., 1998; Pasternak and Greenlee, 2005). This led to the view that working memory information is represented in persistent activity in a brain network that includes the prefrontal cortex (Riley and Constantinidis, 2016). More recently, the use of multivariate approaches demonstrated that the content maintained in working memory does not require stable persistent activity (Wolff et al., 2017), but is encoded in dynamic neural patterns and "activity-silent" states (Meyers et al., 2008; Stokes, 2015). Our results suggest that previously described persistent and dynamic patterns of neural activity may be reflective of two different working memory processes. First, a very stable persistent activity involving the ventrolateral prefrontal cortex is associated with the selection process that prioritizes or reactivates a specific sensory content (**Fig 4**). Then, the sensory content is transiently reactivated upon presentation of the cue (**Fig 5**), consistent with the dynamic population coding identified in primate studies (Meyers et al., 2008; Stokes, 2015).

Conclusions

Our study identified spatiotemporal neural dynamics of the selection and maintenance of a working memory content as it gets manipulated. Evidence is presented in favor of a role for the ventrolateral prefrontal cortex in the selection rather than the maintenance of working memory content, through a stable and frequency-specific neural representation. The neural representation of working memory content was transformed from the initial visual encoding into a different and transiently reactivated memory representation in a posterior brain network. These results may help reconcile different views on the persistent and dynamic features of spatiotemporal neural representations of working memory.

Materials and Methods

Resource sharing

All preprocessing and MEG analysis pipeline are available at https://github.com/romquentin/decod_WM_Selection_and_maintenance and raw MEG will be available soon at https://www.mcgill.ca/bic/resources/omega. Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Romain Quentin (romain.quentin@nih.gov).

Participants and experimental sessions

35 healthy volunteers participated in the study after providing informed consent. They all had normal physical and neurological examinations and normal or corrected-to-normal vision. Participants who reached 75% correct responses during the working memory task in a screening session returned for one structural MRI and two magnetoencephalography (MEG) sessions (23 participants: 17 women, 6 men, mean age = 26.6 ± 6.7). One participant moved out from the area and did only one MEG session.

Visual Working Memory Task

Visual stimuli were displayed using MATLAB (Mathworks, Natick, MA, USA) and the Psychophysics Toolbox (Psychtoolbox-3) (Brainard, 1997) running on a MacBook Pro laptop computer. During the MEG session, visual stimuli were back projected on a translucent screen in front of the participants. Each trial started with the fixation dot in the middle of the screen. Participants were instructed to focus on the fixation dot during the entire trial. After 400ms (± 50ms jitter), two visual gratings, one in each half of the visual field, were simultaneously presented for 100ms (**Fig 1**). Each grating had one out of five possible spatial frequencies (1, 1.5, 2.25, 3.375 or 5.06 cycles/degree) and one out of five possible orientations (-72, -36, 0, 36, 73 in degree, 0 being the vertical). A visual cue, lasting 100ms, was presented 900 ms (± 50ms jitter) after the stimulus onset, indicating the side (spatial rule indicating left or right) and the feature (feature rule indicating orientation or spatial frequency) to be remembered. A probe was provided 1600ms (± 50ms jitter) after the cue onset, and participants had to match the cued item with the probe (similar or different) by responding with their right index and middle finger on a button box. The probe disappeared when the participants gave their response. The fixation dot turned green for a correct answer or red for an incorrect one during 100ms at the end of each trial. Eye movements were monitored across the trial with an eye-tracker (Eyelink 1000, SR Research, Mississauga, ON, Canada) to ensure correct central fixation. Fixation was

considered broken when participants' gaze was recorded outside a circular spot with a 2.5 visual degree radius around the center of the fixation dot or if they blinked during the period from the stimulus onset to the probe onset. In that eventuality, participants received an alert message on the screen and the trial was shuffled with the rest of the remaining trials and repeated. Each session was composed of 400 trials with correct fixation interspersed with rest periods every block of 50 trials. A total of 800 trials with correct fixation were obtained from each participant during 2 MEG sessions (except one who came for only one MEG session, 400 trials). Group average behavioral performance during this task reached 83%. Participants were better at recalling the orientation than the spatial frequency trials (85 *vs*. 81%, $p<0.001$). No difference was found between performance in left and right cue trials (**Supp Fig 1**).

One-back task

In 22 MEG sessions, a one-back task (160 trials with correct fixation) was performed prior to the working memory task to control for the visual processing of the cue. During this task, one of the four cues used in the working memory task appeared every 1500ms and the participant had simply to press a button if two consecutive cues were similar. Eye movement monitoring was performed. If participants broke visual fixation, the trial was shuffled with the remaining trials and repeated. Group average behavioral performance during this task reached 89% of correct response.

MRI acquisition and preprocessing

Magnetic Resonance Imaging (MRI) data were acquired with a Siemens Skyra 3T scanner using a 32-channel coil. High-resolution ($0.93 \times 0.93 \times 0.9$ mm^3) 3D magnetization prepared rapid gradient echo (MPRAGE) T1-weighted images were acquired (repetition time = 1900 ms; echo time = 2.13 ms; matrix size = $256 \times 256 \times 192$). A stereotactic neuronavigation system (Brainsight, Rogue Research, Montreal, QC, Canada) was used before the MEG recordings to record MRI coordinates of the three head position coils placed on the nasion and pre-auricular points. These coil position coordinates were used to co-register the head with the MEG sensors for source reconstruction. Brain surfaces were reconstructed using the FreeSurfer software package (Dale et al., 1999; Fischl et al., 1999). A forward model was generated from the segmented and meshed MRI using Freesurfer (Fischl, 2012) and MNE-python (Gramfort et al., 2013) and co-registered to the MRI coordinates with the head position coils.

MEG recordings

Neuromagnetic activity was recorded with a sampling rate of 1,200 Hz on the NIH 275-channel CTF magnetoencephalography (MEG International Services, Ltd., Coquitlam, BC, Canada). The MEG apparatus was housed in a magnetically shielded room. During recording, participants were seated alone in the shielded MEG room and their head was centrally positioned within the sensor array. The head position was recorded before and after each block. If the difference between the two recordings exceeded 3mm, participants were asked to reposition their head to its original position while their real-time head position was displayed. Brain MEG activity was band-passed filtered in the range of 0.5 to 25 Hz and decimate by 10, resulting in a sampling frequency of 120Hz. MEG signal was epoched at the onset of the stimulus (-0.2s, 0.9s), the onset of the cue (-0.2s, 1.5s) and the onset of the probe (-0.2, 0.4s). The two MEG sessions per participant were concatenated. The epoch data for the three events were all baselined between -0.2 and 0s according to the stimulus onset, except for the source analyses where the cue epoch was baselined between -0.2 and 0s according to the cue onset. A Digital-to-Analog converter was used to record the eye tracker signal with the MEG acquisition system.

MEG Multivariate Pattern Analysis (MVPA)

Data was analyzed with multivariate linear modeling, following King et al's preprocessing pipeline (King and Dehaene, 2014; King et al., 2016) and implemented in MNE-python (Gramfort et al., 2013). MVPA decoding aimed at predicting the value of a specific variable y (for example the cued spatial frequency or line orientation) from the brain signal. The analysis consists of 1) fitting a linear estimator w to a training subset of X (X_{train}), 2) from this estimator, predicting an estimate ($\bar{y}_{test}$) of the variable (y_{test}) on a separate test subset (X_{test}) and finally 3) assessing the decoding score of this prediction as compared to the ground truth ($score(y_{test}, \bar{y}_{test})$). In our analysis, MEG data (X) was whitened by using a standard scaler that z scored each channel at each time point across trials. An *l2* linear model was then fitted to find the hyperplane (w) that maximally predicts the variable of interest (y). All parameters were set to their default values as provided by the Scikit-Learn package (Pedregosa et al., 2011). A logistic regression has been used to decode categorical data (cue side or cue type) and a ridge regression to decode the spatial frequency. A combination of two ridge regressions was used to perform circular correlations to decode the orientation, fitted to predict $\sin(y)$ and $\cos(y)$. The predicted angle ($\bar{y}$) was estimated from the arctangent of the resulting sine and cosine: $\bar{y} = artan2(\bar{y}_{sin}, \bar{y}_{cos})$. Each estimator was fitted on each participant separately, across all MEG

sensors (or sources) and at a unique time sample (sampling frequency = 120Hz). The cross-validation was performed using a 12-fold stratified folding, such that each estimator was trained on 11/12th of the trials (training set) and then generated a prediction on the remaining 1/12th trials (testing set). Ordinal effects (decoding of spatial frequency) were summarized with a Spearman Correlation R coefficient (range between -1 and 1 with chance = 0). Categorical effects (decoding of cue side and cue type) were summarized with the area under the curve (AUC) (range between 0 and 1 with chance = 0.5). Circular decoding was summarized by computing the mean absolute difference between the predicted angle ($\bar{y}$) and the true angle (y) (range between 0 and π, chance = π/2). To facilitate visualizations, this "error" metric was transformed into an "accuracy" metric (range between –π/2 and π/2, chance = 0) (King et al., 2016). In addition, within each analysis, the temporal generalization was computed. Each estimator trained on time *t* was tested on its ability to predict a given trial at time *t'*, in order to estimate the similarity of the coding pattern at *t* and *t'* and thus the stability of the neural representation. Results of this temporal generalization are presented in a 2D matrix with training time on the vertical axis and testing time on the horizontal axis. All decoding analyses were performed with the MNE-python (Gramfort et al., 2013) and Scikit-Learn packages (Pedregosa et al., 2011). To test the similarity between the neural representation during visual perception and working memory, estimators were either trained on stimulus decoding and test on the memory content or the inverse. Estimators were trained separately for trials with left and right cue. An estimator trained/tested on the left spatial frequency was trained/tested on the cued spatial frequency only when the cue indicated the left side of the stimulus and the same for the right. Results of this generalization across condition were then averaged between left and right for statistical test and visualization. This work utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov).

MEG source reconstruction

To estimate the time series in source space, the Linearly Constrained Minimum Variance (LCMV) beamformer method was computed on single trial data using MNE-python (Gramfort et al., 2013). The regularized noise covariance matrix was computed on a pre-stimulus period (-0.3, 0s according to stimulus onset). The regularized data covariance was computed during a period starting 40ms after the event of interest (either stimulus, cue or probe onset) until the end of each epoch (respectively 900ms, 1500ms and 400ms). MVPA analysis was applied on this single trial source time series. To investigate the spatial distribution of brain regions contributing to the decoding performance, and because raw classifier weights are difficult to interpret, we

transformed these weights into patterns using the recently described method of Haufe et al (2014). For each analysis, these individual patterns were morphed and averaged on the surface-based "fsaverage" template of Freesurfer (Fischl, 2012). Then, a principal component analysis was applied and the two first components were plotted on the inflated average brain.

Statistical Analysis

Each analysis was first performed within each subject separately using all meaningful trials, i.e., all trials were used to decode visual attributes of the stimulus or the probe, cue side and cue type, and trials with a cue indicating either the spatial frequency or the orientation were used to decode the specific memory content. Statistical analyses were based on second-level tests across participant and were performed on the temporal generalization or time frequency matrix of decoding performance with a non-parametric one sample t-test corrected for multiple comparisons with cluster-based permutations (Maris and Oostenveld, 2007), using the default parameters of the MNE-python *spatio_temporal_cluster_1samp_test* function. Color-filled areas on decoding performance curves or dashed contour on temporal generalization and time frequency matrices correspond to p-value < 0.05 resulting from this permutation test. To test the decoding performance on a large window, decoding performances were averaged across all time samples in each participant and epoch period starting from the event onset (either stimulus, cue or probe) and then tested at the group level with a one sample t-test against chance level (***, **, * indicate respectively $p < 0.001$, $p < 0.01$ and $p < 0.05$).

Bibliography

Baddeley, A. (2010). Working memory. Curr. Biol. *20*, R136–R140.
Baddeley, A.D., and Hitch, G. (1974). Working Memory. Psychol. Learn. Motiv. *8*, 47–89.
Bauer, R.H., and Fuster, J.M. (1976). Delayed-matching and delayed-response deficit from cooling dorsolateral prefrontal cortex in monkeys. J. Comp. Physiol. Psychol. *90*, 293–302.
Brainard, D.H. (1997). The Psychophysics Toolbox. Spat. Vis. *10*, 433–436.
Buschman, T.J., and Miller, E.K. (2007). Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. Science *315*, 1860–1862.
Buschman, T.J., Denovellis, E.L., Diogo, C., Bullock, D., and Miller, E.K. (2012). Synchronous oscillatory neural ensembles for rules in the prefrontal cortex. Neuron *76*, 838–846.
Chelazzi, L., Duncan, J., Miller, E.K., and Desimone, R. (1998). Responses of neurons in inferior temporal cortex during memory-guided visual search. J. Neurophysiol. *80*, 2918–2940.
Christophel, T.B., Hebart, M.N., and Haynes, J.-D. (2012). Decoding the contents of visual short-term memory from human visual and parietal cortex. J. Neurosci. Off. J. Soc. Neurosci. *32*, 12983–12989.
Christophel, T.B., Klink, P.C., Spitzer, B., Roelfsema, P.R., and Haynes, J.-D. (2017). The Distributed Nature of Working Memory. Trends Cogn. Sci. *21*, 111–124.
Cichy, R.M., Ramirez, F.M., and Pantazis, D. (2015). Can visual information encoded in cortical columns be decoded from magnetoencephalography data in humans? NeuroImage *121*, 193–204.
Courtney, S.M., Petit, L., Maisog, J.M., Ungerleider, L.G., and Haxby, J.V. (1998). An area specialized for spatial working memory in human frontal cortex. Science *279*, 1347–1351.
Curtis, C.E., and D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. Trends Cogn. Sci. *7*, 415–423.
Dale, A.M., Fischl, B., and Sereno, M.I. (1999). Cortical surface-based analysis. I. Segmentation and surface reconstruction. NeuroImage *9*, 179–194.
D'Esposito, M., and Postle, B.R. (2015). The cognitive neuroscience of working memory. Annu. Rev. Psychol. *66*, 115–142.
Emrich, S.M., Riggall, A.C., Larocque, J.J., and Postle, B.R. (2013). Distributed patterns of activity in sensory cortex reflect the precision of multiple items maintained in visual short-term memory. J. Neurosci. Off. J. Soc. Neurosci. *33*, 6516–6523.
Ester, E.F., Sprague, T.C., and Serences, J.T. (2015). Parietal and Frontal Cortex Encode Stimulus-Specific Mnemonic Representations during Visual Working Memory. Neuron *87*, 893–905.
Fischl, B. (2012). FreeSurfer. NeuroImage *62*, 774–781.
Fischl, B., Sereno, M.I., Tootell, R.B., and Dale, A.M. (1999). High-resolution intersubject averaging and a coordinate system for the cortical surface. Hum. Brain Mapp. *8*, 272–284.
Fuentemilla, L., Penny, W.D., Cashdollar, N., Bunzeck, N., and Düzel, E. (2010). Theta-Coupled Periodic Replay in Working Memory. Curr. Biol. *20*, 606–612.
Funahashi, S., and Kubota, K. (1994). Working memory and prefrontal cortex. Neurosci. Res. *21*, 1–11.
Funahashi, S., Bruce, C.J., and Goldman-Rakic, P.S. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. J. Neurophysiol. *61*, 331–349.
Fuster, J.M., and Alexander, G.E. (1971). Neuron activity related to short-term memory. Science *173*, 652–654.
Goldman-Rakic, P.S. (1995). Cellular basis of working memory. Neuron *14*, 477–485.

Gramfort, A., Luessi, M., Larson, E., Engemann, D.A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., et al. (2013). MEG and EEG data analysis with MNE-Python. Brain Imaging Methods *7*, 267.
Griffin, I.C., and Nobre, A.C. (2003). Orienting Attention to Locations in Internal Representations. J. Cogn. Neurosci. *15*, 1176–1194.
Han, X., Berg, A.C., Oh, H., Samaras, D., and Leung, H.-C. (2013). Multi-voxel pattern analysis of selective representation of visual working memory in ventral temporal and occipital regions. NeuroImage *73*, 8–15.
Harrison, S.A., and Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. Nature *458*, 632.
Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., and Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. NeuroImage *Complete*, 96–110.
Jacobsen, C.F. (1935). Functions of frontal association area in primates. Arch. Neurol. Psychiatry *33*, 558–569.
King, J.-R., and Dehaene, S. (2014). Characterizing the dynamics of mental representations: the temporal generalization method. Trends Cogn. Sci. *18*, 203–210.
King, J.-R., Pescetelli, N., and Dehaene, S. (2016). Brain Mechanisms Underlying the Brief Maintenance of Seen and Unseen Sensory Information. Neuron *92*, 1122–1134.
Klingberg, T. (2010). Training and plasticity of working memory. Trends Cogn. Sci. *14*, 317–324.
Kumar, S., Joseph, S., Gander, P.E., Barascud, N., Halpern, A.R., and Griffiths, T.D. (2016). A Brain System for Auditory Working Memory. J. Neurosci. *36*, 4492–4505.
Lee, S.-H., and Baker, C.I. (2016). Multi-Voxel Decoding and the Topography of Maintained Information During Visual Working Memory. Front. Syst. Neurosci. *10*.
Lee, S.-H., Kravitz, D.J., and Baker, C.I. (2013). Goal-dependent dissociation of visual and prefrontal cortices during working memory. Nat. Neurosci. *16*, 997.
Lewis-Peacock, J.A., Drysdale, A.T., Oberauer, K., and Postle, B.R. (2012). Neural evidence for a distinction between short-term memory and the focus of attention. J. Cogn. Neurosci. *24*, 61–79.
Lundqvist, M., Rose, J., Herman, P., Brincat, S.L., Buschman, T.J., and Miller, E.K. (2016). Gamma and Beta Bursts Underlie Working Memory. Neuron *90*, 152–164.
Makovski, T., Sussman, R., and Jiang, Y.V. (2008). Orienting attention in visual working memory reduces interference from memory probes. J. Exp. Psychol. Learn. Mem. Cogn. *34*, 369–380.
Maris, E., and Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. J. Neurosci. Methods *164*, 177–190.
Meyers, E.M., Freedman, D.J., Kreiman, G., Miller, E.K., and Poggio, T. (2008). Dynamic Population Coding of Category Information in Inferior Temporal and Prefrontal Cortex. J. Neurophysiol. *100*, 1407–1419.
Mostert, P., Albers, A.M., Brinkman, L., Todorova, L., Kok, P., and Lange, F.P. de (2017). Eye movement-related confounds in neural decoding of visual working memory representations. BioRxiv 215509.
Murray, A.M., Nobre, A.C., Clark, I.A., Cravo, A.M., and Stokes, M.G. (2013). Attention Restores Discrete Items to Visual Short-Term Memory. Psychol. Sci. *24*, 550–556.
Murray, J.D., Bernacchia, A., Roy, N.A., Constantinidis, C., Romo, R., and Wang, X.-J. (2017). Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. Proc. Natl. Acad. Sci. U. S. A. *114*, 394–399.
Myers, N.E., Stokes, M.G., and Nobre, A.C. (2017). Prioritizing Information during Working Memory: Beyond Sustained Internal Attention. Trends Cogn. Sci. *21*, 449–461.

Pasternak, T., and Greenlee, M.W. (2005). Working memory in primate sensory systems. Nat. Rev. Neurosci. *6*, 97–107.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.
Petrides, M. (1996). Specialized systems for the processing of mnemonic information within the primate frontal cortex. Philos. Trans. R. Soc. Lond. B. Biol. Sci. *351*, 1455-1461; discussion 1461-1462.
Petrides, M. (2005). Lateral prefrontal cortex: architectonic and functional organization. Philos. Trans. R. Soc. B Biol. Sci. *360*, 781–795.
Phillips, S., and Takeda, Y. (2009). Greater frontal-parietal synchrony at low gamma-band frequencies for inefficient than efficient visual search in human EEG. Int. J. Psychophysiol. *73*, 350–354.
Pollmann, S., and von Cramon, D.Y. (2000). Object working memory and visuospatial processing: functional neuroanatomy analyzed by event-related fMRI. Exp. Brain Res. *133*, 12–22.
Reverberi, C., Görgen, K., and Haynes, J.-D. (2012). Compositionality of rule representations in human prefrontal cortex. Cereb. Cortex N. Y. N 1991 *22*, 1237–1246.
Riggall, A.C., and Postle, B.R. (2012). The relationship between working memory storage and elevated activity as measured with functional magnetic resonance imaging. J. Neurosci. Off. J. Soc. Neurosci. *32*, 12990–12998.
Riley, M.R., and Constantinidis, C. (2016). Role of Prefrontal Persistent Activity in Working Memory. Front. Syst. Neurosci. *9*.
Rose, N.S., LaRocque, J.J., Riggall, A.C., Gosseries, O., Starrett, M.J., Meyering, E.E., and Postle, B.R. (2016). Reactivation of latent working memories with transcranial magnetic stimulation. Science *354*, 1136–1139.
Sauseng, P., Klimesch, W., Stadler, W., Schabus, M., Doppelmayr, M., Hanslmayr, S., Gruber, W.R., and Birbaumer, N. (2005). A shift of visual spatial attention is selectively associated with human EEG alpha activity. Eur. J. Neurosci. *22*, 2917–2926.
Serences, J.T., Ester, E.F., Vogel, E.K., and Awh, E. (2009). Stimulus-specific delay activity in human primary visual cortex. Psychol. Sci. *20*, 207–214.
Sreenivasan, K.K., Curtis, C.E., and D'Esposito, M. (2014). Revisiting the role of persistent neural activity during working memory. Trends Cogn. Sci. *18*, 82–89.
Stokes, M.G. (2015). 'Activity-silent' working memory in prefrontal cortex: a dynamic coding framework. Trends Cogn. Sci. *19*, 394–405.
Stokes, M.G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., and Duncan, J. (2013). Dynamic Coding for Cognitive Control in Prefrontal Cortex. Neuron *78*, 364–375.
Stokes, M.G., Wolff, M.J., and Spaak, E. (2015). Decoding Rich Spatial Information with High Temporal Resolution. Trends Cogn. Sci. *19*, 636–638.
Sugase-Miyamoto, Y., Liu, Z., Wiener, M.C., Optican, L.M., and Richmond, B.J. (2008). Short-Term Memory Trace in Rapidly Adapting Synapses of Inferior Temporal Cortex. PLOS Comput. Biol. *4*, e1000073.
Vidaurre, D., Myers, N., Stokes, M., Nobre, A.C., and Woolrich, M.W. (2018). Temporally unconstrained decoding reveals consistent but time-varying stages of stimulus processing. BioRxiv 260943.
Vogel, E.K., McCollough, A.W., and Machizawa, M.G. (2005). Neural measures reveal individual differences in controlling access to working memory. Nature *438*, 500–503.
Warden, M.R., and Miller, E.K. (2010). Task-Dependent Changes in Short-Term Memory in the Prefrontal Cortex. J. Neurosci. *30*, 15801–15810.

Wolff, M.J., Jochim, J., Akyürek, E.G., and Stokes, M.G. (2017). Dynamic hidden states underlying working-memory-guided behavior. Nat. Neurosci. *20*, 864–871.
Woolgar, A., Thompson, R., Bor, D., and Duncan, J. (2011). Multi-voxel coding of stimuli, rules, and responses in human frontoparietal cortex. NeuroImage *56*, 744–752.
Worden, M.S., Foxe, J.J., Wang, N., and Simpson, G.V. (2000). Anticipatory biasing of visuospatial attention indexed by retinotopically specific alpha-band electroencephalography increases over occipital cortex. J. Neurosci. Off. J. Soc. Neurosci. *20*, RC63.

Acknowledgements

The intramural NINDS, DIR, NIH as well as the NIMH MEG and FMRIF facilities on the NIH campus in Bethesda (MD) contributed to this research. This project received funding from the FYSSEN foundation, the Bettencourt-Schueller Foundation and the Philippe Foundation. We thank Dr. R. Coppola and Dr. T. Holroyd for their help in managing the NIH MEG facility and technical advice and Dr. C. Baker and Dr. M. Vernet for their advice on the manuscript. We are grateful to the MNE and Scikit-Learn communities for their very precious help and support.

Author contributions

Conceptualization, R.Q and J.R.K. Methodology, R.Q and J.R.K. Software, J.R.K. Formal analysis, R.Q. Investigation. R.Q, E.S, N.F, R.T and E.B. Writing – Original Draft, R.Q. and L.G.C. Writing – Review & Editing. R.Q, L.G.C, J.R.K, E.B, E.S, R.T. Supervision. L.G.C. Funding Acquisition. R.Q, L.G.C.