Expert programmers have fine-tuned cortical representations of source code

Yoshiharu Ikutani[1], Takatomi Kubo[1,*], Satoshi Nishida[2], Hideaki Hata[1], Kenichi Matsumoto[1], Kazushi Ikeda[1], and Shinji Nishimoto[2]

[1]Graduate School of Science and Technology, Division of Information Science, Nara Institute of Science and Technology (NAIST), Ikoma, Nara 630-0192, Japan
[2]Center for Information and Neural Networks (CiNet), National Institute of Information and Communications Technology (NICT), Suita, Osaka 565-0871, Japan
[*]takatomi-k@is.naist.jp

ABSTRACT

Expertise enables humans to achieve outstanding performance on domain-specific tasks, and programming is no exception. Many have shown that expert programmers exhibit remarkable differences from novices in behavioral performance, knowledge structure, and selective attention. However, the underlying differences in the brain are still unclear. We here address this issue by associating the cortical representation of source code with individual programming expertise using a data-driven decoding approach. This approach enabled us to identify seven brain regions, widely distributed in the frontal, parietal, and temporal cortices, that have a tight relationship with programming expertise. In these brain regions, functional categories of source code could be decoded from brain activity and the decoding accuracies were significantly correlated with individual behavioral performances on source-code categorization. Our results suggest that programming expertise is built up on fine-tuned cortical representations specialized for the domain of programming.

Introduction

Programming expertise is one of the most notable capabilities in the current computerized world. Since human software developers keep playing a central role in every software project and directly impact its success, this relatively new type of expertise is attracting increasing attention from modern industries[1,2] and educations[3,4]. Moreover, software engineering researchers repeatedly found huge variations in productivity even between programmers with the same level of experience[5–7]. Several previous studies showed the psychological characteristics of expert programmers in their behaviors[8,9], knowledge structures[10–12], and eye movements[13–16]. Although these studies clearly illustrate the behavioral specificity of expert programmers, it remains unclear what neural bases differentiate expert programmers from novices.

Recent studies have investigated the brain activity of programmers using functional magnetic resonance imaging (fMRI) to examine their cognitive mechanisms. Siegmund *et al.* contrasted brain activity during program output estimations against syntax error searches and showed that the processes of program output estimations activated left-lateralized brain regions in the middle temporal gyrus, inferior parietal lobule, middle and inferior frontal gyri[17,18]. Several studies have also tried to investigate neural correlates of subject-wise programming expertise but failed to find a systematic trend. For example, a study reported no significant correlation between BOLD activation strength and subjective estimates of programming experience[19]. Although an exploratory study argued the correlation between activity pattern discriminability and students' GPA score[20], the assumed relationship of GPA scores to programming expertise was ambiguous and not empirically validated. Further, the main limitation of these prior studies is the use of a single homogeneous subject group that only covered a small range of programming expertise. Recruitment of more diverse subjects in terms of their programming expertise may enable us to elucidate the potential differences of brain functions related to the expertise[21].

In the present study, we aim to identify the neural bases of programming expertise that contribute outstanding performances of expert programmers and provide a clue to describe how the brain accommodates such behavioral superiority in programming. To do this, we defined two fundamental factors in our experimental design: An objective indicator of programming expertise and a laboratory task that efficiently exhibits experts' superior performances under the general constraints of fMRI experiments. For the first factor, we adopted programmers' ratings in competitive programming contests (AtCoder), which are objectively determined by the relative positions of their actual performances among thousands of programmers[22]. We recruited top- and middle-rated programmers as well as novice controls to cover a wide range of programming expertise in our fMRI experiment (Table.1). For the second factor, we developed the program categorization task and confirmed that behavioral performances of

this task were significantly correlated with the programming expertise indicator. This confirmation allows us to expect the tight association between individual programming expertise and the brain activity patterns recorded by fMRI while subjects performed this laboratory task.

To examine the brain activity patterns underlying expert programmers' behavioral superiority, we employ a decoding framework that learns the relationship between multi-voxel activity patterns in the brain and functional categories of source code. This framework was motivated by prior studies that contrasted multi-voxel activity patterns of experts against novices and demonstrated that domain-specific expertise generally associates with representational changes in the brain[23–26]. Here we hypothesized that higher programming expertise relates to specific multi-voxel pattern representations, potentially influenced by their domain-specific knowledge and training experiences. In our experiment, we presented subjects with a set of Java code snippets implementing eleven fundamental algorithms categorized into four functional categories. A support vector machine classifier (decoder) was then trained and tested to predict functional categories of code snippets from brain activity measured by fMRI while programmers categorized the code snippets. To explore the potential loci of programming expertise in a data-driven manner, we adopted whole-brain searchlight analysis[27].

We demonstrate that functional categories of program source code can be decoded from programmers' brain activity and decoding accuracies in seven distinct brain regions are significantly correlated with individual behavioral performances that reflect programming expertise. Furthermore, we show that decoding accuracies of subordinate-level categories on two brain regions are significantly correlated with individual behavioral performance, even though such discriminations are not explicitly required by the tasks. These results suggest that expert programmers' outstanding performances depend on fine-tuned cortical representations of source code and such cortical representation refinements might be related to the acquisition of advanced-level programming expertise.

Results

Behavioral performance on the program categorization task.

We first evaluated behavioral performance on the program categorization task, the laboratory task designed for capturing expert programmers' domain-specific knowledge used in our experiment. We collected 72 Java code snippets from open codeset provided by AIZU ONLINE JUDGE, each implementing one of the eleven fundamental algorithms categorized into four functional categories (Fig.1a; see Supplementary Table 1 and 2 for detailed descriptions of each category and subcategory). In the experiment, subjects were presented with the code snippets and asked to categorize them into one of four functional category classes while fMRI signals were measured (Fig.1b). Every subject performed the program categorization task 216 times (36 trials $\times$ 6 separate runs) inside the MRI scanner and 72 code snippets were each presented three times (see Supplementary Figure 1 and Table 3 for examples and statistics of the code snippets). Subjects responded via pressing buttons placed under the right hand to indicate which class was most plausible for each code snippet and all response data were automatically collected for the calculation of individual behavioral performance. Additionally, behavioral performances on the *subcategory* categorization were assessed by the post-MRI experiments conducted within ten days after the fMRI measurements. Note that the existence of *subcategory* classes had never been revealed until the end of the fMRI experiments to make subjects concentrate on the *category* categorizations inside the scanner.

As a result, we confirmed significant correlations between behavioral performances on the program categorization task and the programming expertise indicator i.e. AtCoder rate, which quantified the relative positions of their actual performances in competitive programming contests. We observed a positive correlation between AtCoder rate (M = 954.3, SD = 864.6) and behavioral performance (M = 76.0, SD = 13.5 [%]), r = 0.722, p = 0.000007, n = 30 (Fig.2a). This positive correlation was remained if we exclude non-rate-holder subjects to avoid regarding them as zero-rated subjects; r = 0.593, p = 0.0059, n = 20. As shown in Fig.2b, we additionally found a positive correlation between AtCoder rate and behavioral performance on *subcategory* categorization (M = 65.9, SD = 17.0 [%]), r = 0.735, p = 0.000004, n = 30. This significant correlation was kept even if we exclude non-rate-holder subjects; r = 0.688, p = 0.0008, n = 20. From all behavioral data, we certainly concluded that behavioral performances on the program categorization task significantly correlated with programming expertise. The behavioral evidence allowed us to expect that individual programming expertise was reflected in the brain activity patterns measured using fMRI while subjects performed this laboratory task.

Mutli-voxel activity patterns associated with programming expertise.

Our decoding framework aims to learn the relationship between programmers' brain activity patterns and functional categories of source code to identify the cortical representations that associate with programming expertise (Fig.3). We employed a whole-brain searchlight analysis[27] as a data-driven approach. A four-voxel-radius sphered searchlight, covering 251 voxels at once, was systematically shifted throughout the brain and decoding accuracy was quantified on each searchlight location. We used a linear-kernel support vector machine (SVM) classifier and calculated decoding accuracy as a ratio of correct-classifications out of all classifications (chance-level accuracy = 25%). The SVM classifier was trained and tested for each subject independently

using a one-run-out cross-validation procedure, which iteratively treated data in a single run for test and others for training, and the value of decoding accuracy was finally determined by averaging results of all iterations. In theory, decoding accuracy would be higher when a target population of voxels clearly differentiates functional categories of program source code.

We first examined where we could decode the functional categories of source code from programmers' brain activity. Fig.4 visualizes the searchlight centers that showed significantly high decoding accuracy than chance estimated from all subject data using a relatively strict whole-brain statistical threshold (voxel-level $p < 0.05$ FWE-corrected). The figure indicates that significant decoding accuracies were observed in the broad areas of bilateral occipital cortices, parietal cortices, posterior and ventral temporal cortices, as well as the bilateral frontal cortices around inferior frontal gyri. Given the result, we certainly confirmed that functional categories of source code were represented in the widely distributed brain areas and the cortical representations of each *category* class were linearly separable by a simple SVM classifier. Note that, in this decoding analysis, we only examined that "where" significant decoding accuracies exist; not judge whether or not these cortical representations were correlated with individual programming expertise.

To associate the cortical representation of source code with individual programming expertise, we investigated a linear correlation between behavioral performances and decoding accuracies for each searchlight location. Fig.5a visualizes the searchlight centers that showed significantly high correlation coefficients using thresholds of voxel-level $p < 0.001$ uncorrected and cluster-level $p < 0.05$ FWE-corrected. We observed significant correlations in the areas of bilateral inferior frontal gyri pars triangularis (IFG Tri), right superior frontal gyrus (SFG), left inferior parietal lobule (IPL), left middle and inferior temporal gyrus (MTG / IT); see the slice-width visualization shown as Fig.5b and Supplementary Table 4 for the list of significant clusters. In this correlation analysis, the right IFG Tri showed the highest peak correlation coefficient ($r = 0.79$, $p < 10^{-6}$ uncorrected, Fig.5c). These results provided evidence that cortical representations in the distinct brain areas mainly located in frontal, parietal, and temporal cortices were significantly associated with experts' outstanding performances on the program categorization task. In contrast, cortical representations in the bilateral occipital cortices including early visual areas did not show a significant correlation to individual behavioral performances, while significant decoding accuracies were broadly observed in the cortices shown as Fig.4.

Previous two analyses separately showed where significant decoding accuracies exist and whether these decoding accuracies significantly correlate with behavioral performances. To achieve more validated evidence for the cortical representations associated with programming expertise, we integrated these two analyses and identified searchlight centers that had sufficient information to represent functional categories of source code and their decoding accuracies significantly correlated with individual behavioral performance. As a result, we found 1,205 searchlight centers (equal to 0.79%) that survived from both statistical thresholds of decoding accuracy and correlation to behavioral performances; shown as red-colored dots in Fig.6a. As shown in Fig.6b, the survived searchlight centers were mainly observed in the bilateral IFG Tri, left IPL, left supramarginal gyrus (SMG), left MTG/IT, and right middle frontal gyrus (MFG). The complementary *sensitivity* analysis[28] using a five-voxel-radius searchlight showed the almost same tendency, indicating that the results were not limited to a specific searchlight radius parameter (see Supplementary Figure 3 and Supplementary Table 5). Since we have demonstrated that individual behavioral performances were significantly correlated with the expertise indicator in competitive programming contests (Fig.2a), this result revealed a tight association between high-level programming expertise and the improvement of decoding accuracy in these seven brain regions.

Cortical representations of subcategory information.

We next investigated where we could decode the *subcategory* of source code from programmers' brain to examine finer-level cortical representations. In our experiment, subjects responded 'sort' when he/she has been presented with the code snippets implementing one of three different sorting algorithms; i.e. bubble, insertion, and selection sorts (Fig.1a). This cognitive process could be considered as a generalization process that incorporates different but similar algorithms (*subcategory*) into a more general functionality class (*category*). Additionally, several psychologists indicated that experts specifically show high performances in subordinate-level categorizations as well as basic-level categorizations[29]. In fact, we have observed that the ability to differentiate *subcategory* classes significantly correlated to the programming expertise indicator in competitive programming (Fig.2b). This evidence implies that programmers' brain activity patterns may automatically respond to the detailed functional difference of source code. The decoding accuracy of *subcategory* may be correlated with programming expertise, even though they classified only *category* classes, not *subcategory*, of given code snippets and the existence of *subcategory* classes had never been revealed until the end of fMRI experiment.

We employed searchlight analysis with the same setting as used in the previous analysis to reveal the spatial distribution of significant *subcategory* decoding accuracies and significant correlations to behavioral performances. Fig.7 illustrated the searchlight centers that showed significantly high *subcategory* decoding accuracy than chance (9.72%; corrected for imbalanced exemplars) using a threshold of voxel-level $p < 0.05$ FWE-corrected. The figure indicated that the extent of significant *subcategory* decoding accuracies was similar to those of the *category* decoding result shown in Fig.4. Linear correlation

between *subcategory* decoding accuracies and individual behavioral performances was then assessed using thresholds of voxel-level $p < 0.001$ uncorrected and cluster-level $p < 0.05$ FWE-corrected (Fig.8). As a result, only a cluster on the left SMG and superior temporal gyrus (STG) showed a significant correlation; the peak correlation coefficient was observed in the left STG ($r = 0.72$, $p < 10^{-5}$ uncorrected, Fig.8c). Finally, we integrated the results from decoding and correlation analysis of *subcategory* and confirmed that 120 searchlight centers (equal to 0.08%) on the left SMG and STG survived from both statistical thresholds of decoding accuracy and correlation to behavioral performances; shown as red-colored dots in Fig.9a. The complementary *sensitivity* analysis using a five-voxel-radius searchlight indicated that these results were consistently observed across the two searchlight radius parameters (see Supplementary Figure 4 and Supplementary Table 6). These results suggest that cortical representations of fine functional categories on the left SMG and STG may play an important role in achieving advanced-level programming expertise, even though the representations are not explicitly required by the tasks.

Discussion

We have shown that functional categories of source code can be decoded from programmers' brain activity measured using fMRI and decoding accuracies on the bilateral IFG Tri, left IPL, left SMG, left MTG, left IT, and right MFG were significantly correlated with individual behavioral performances on the program categorization task (Fig.6). Furthermore, decoding accuracies of *subcategory* on the left SMG and STG were also strongly correlated with the behavioral performances (Fig.9) while the subordinate-level representations were not directly induced by the performing tasks. Since we have demonstrated in the beginning of this study that the behavioral performances were correlated with the expertise indicator in competitive programming contests (Fig.2), our results revealed a tight association between advanced-level programming expertise and domain-specific cortical representations in these brain areas widely distributed in the frontal, parietal, and temporal cortices.

Previous fMRI studies on programmers have aimed at characterizing how programming-related activities, such as program comprehension and bug detection, take place in the brain[17–20,30,31]. Exceptionally, an exploratory study reported that BOLD signal discriminability between code and text comprehensions was negatively correlated with participants' GPA scores in a university[20]. However, the relationship between GPA scores and programming expertise was ambiguous and the observed correlation was relatively small ($r = -0.44$, $p = 0.016$, $n = 29$). Our aim in the present study was substantially different: We sought the neural bases of programming expertise that contribute expert programmers' outstanding performances. To address the goal, we adopted an objective indicator of programming expertise and recruited a population of subjects covering wide range of programming expertise. It is worth noting that the expertise indicator and behavioral/neural data obtained in this study were completely independent from each other. Because our novel laboratory task well bridged between them, we succeeded to associate programming expertise with programmers' cortical representations in a reasonable way.

Despite the difference in research aims, a subset of brain regions specified in this study was similar to those specified by prior fMRI studies on programmers[17–19]. In particular, this study associated the left IFG, MTG, IPL, SMG with programming expertise while previous studies related them with program comprehension processes. This commonality is remarkable because these results jointly suggest that both program comprehension processes and its related expertise may depend on the same set of brain regions. Providing interpretations of their potential roles in programming expertise would be beneficial for orienting future researches. First, the left IFG Tri and the left posterior MTG are frequently involved in semantic selecting and retrieving tasks[32–35]. Several studies indicated that these two regions are sensitive to cognitive demands for directing semantic knowledge retrieval in a goal-oriented way[36–38]. The involvements of the two regions in our findings may be induced by similar demands specialized for retrieval of program functional category and suggest that higher programming expertise is related to greater abilities of goal-oriented knowledge retrieval. Second, many neuroscientists have shown the left IPL and SMG to be functionally related to visual word reading[39–41] and episodic memory retrieval[42–44]. Both cognitive functions potentially relate to the program categorization task used in our experiment. Visual word reading can be naturally engaged since source code is comprised of many English-like words and subjects may have actively recollected previously-acquired memories to compensate for insufficient clues because they had only ten seconds to categorize the given code snippet. The involvements of the left IPL and SMG in programming expertise suggest that expert programmers might possess different reading strategies and/or depend more on domain-specific memory retrieval than novices.

Other novel findings in the present study were potential involvement of the left IT, right MFG, and right IFG Tri with programming expertise. Importantly, these regions were not specified by previous studies focusing on the relationship between brain activity and program comprehension processes[17–20], suggesting that the regions might be more related to programming expertise than program comprehension processes. Because the left IT is well known for the function in high-level visual processing including word recognition and categorical object representations[45–47], our results may suggest that high-level visual cortex in expert programmers' brain could be fine-tuned by their training experience to realize faster program comprehension process. The right MFG and IFG Tri are functionally related to stimulus-driven attention control[48,49]. The involvement of these two regions suggests that programmers with high-level programming expertise may employ different attention strategies than less-skilled ones. Moreover, additional engagements of right hemisphere regions in experts are common across expertise studies.

For example, chess experts[50] and abacus experts[51,52] showed additional right hemisphere region involvements when performing their domain-specific tasks. Several fMRI studies further suggested that such activation shifts from left to right hemisphere may be related to experts' cognitive strategy changes[50,53]. Cognitive strategy changes have been observed repeatedly in comparisons between expert and novice programmers: A major characteristic is a transition from bottom-up (or textual-driven) to top-down (or goal-driven) program comprehension, which becomes feasible by experts' domain-specific knowledge[9,11,12]. The involvement of the right MFG and IFG Tri observed in this study might be related to such cognitive strategy differences between programmers in the program categorization task.

Our results associated programming expertise with decoding accuracies of not only *category* but also *subcategory*, even though the subordinate-level categorizations were not explicitly required by the performing task. We observed that individual behavioral performances were significantly correlated with *subcategory* decoding accuracies on the left STG and SMG (Fig.9). These two regions are functionally related to pre-lexical and phonological processing in natural language comprehension[32,54,55]. Interestingly, we also found a significant correlation between behavioral performances and *category* decoding accuracies on the temporal regions (left MTG and IT) associated with more semantical processing[35,36,38]. If these functional interpretations could be adaptable to program comprehension processes, it would be intuitive that subordinate concrete concepts (i.e. *subcategory*) of source code are processed in the left STG/SMG and more semantically abstract concepts (i.e. *category*) are represented in the left MTG/IT. This might suggest a hypothesis that expert programmers' brain has a hierarchical semantic processing system to obtain mental representations of source code for multiple levels of abstraction.

The results obtained via the present study were limited to a specific type of programming expertise evaluated by the expertise indicator and laboratory task used in the experiment. We particularly examined the ability to semantically categorize source code that correlated with programming expertise to win high scores in competitive programming contests. The ability to write efficient SQL programs, for example, may be an explicit indicator of another type of programming expertise but this study did not cover. Thus, our results should not be taken to imply the relationship between the neural correlates revealed here and other types of programming expertise that could not be examined by this experiment. However, it is also a fact that we cannot investigate the neural bases of programming expertise without a clear definition of expertise indicator and laboratory task that well fit the general constraints of fMRI experiments. To mitigate the potentially inevitable effects caused by this limitation, we adopted the objective indicator of programming expertise that directly reflects programmers' actual performances and recruited a population of subjects covering a wide range of programming expertise. This study can be a baseline example for future researches to investigate the neural bases of programming expertise and other related abilities.

Our decoding framework specialized for the functional category of source code could be extended by the recent advances of decoding/encoding approaches in combination with distributed feature vectors[56]. Several researchers have demonstrated frameworks to decode arbitrary objects using a set of computational visual futures representing categories of target objects[57] and to decode perceptual experiences evoked by natural movies using word-based distributed representations[58]. Other studies have also used word-based distributed representations to systematically map semantic selectivity across the cortex[59,60]. Meanwhile, researchers in the program analysis domain have proposed distributed representations of source code based on abstract syntax tree (AST)[61,62]. Alon et al., for instance, have presented continuous distributed vectors representing the functionality of source code using AST and path-attention neural network[63]. The combination of recent decoding/encoding approaches and distributed representations of source code may enable us to build a computational model of program comprehension that connecting semantic features of source code to programmers' perceptual experiences.

Methods

Subjects.

Thirty healthy subjects (two females, aged between 20 and 24 years) with normal or corrected-to-normal vision participated in the experiment. All were right-handed (assessed by the Edinburgh Handedness Inventory[64], laterality quotient [LQ] = 83.6 ± 24.0, ranged between +5.9 and +100) and understood basic Java grammars with at least half of year experience on Java programming. We recruited top- and middle-rated programmers as well as novice controls to cover a wide range of programming expertise using programmers' rate in competitive programming contests, which are objectively determined by the relative positions of their actual performances among thousands of programmers. We defined three recruiting criteria in advance of starting to recruit candidate subjects. The criteria recruited top 20% and 21-50% rankers in AtCoder (https://atcoder.jp/) as *Expert* and *Middle* based on the ranking at July 1 2017 and subjects with four years or less programming experience and no experience on competitive programming as *Novice*. Table.1 summarized the detailed demographic information of all recruited subjects. The averaged AtCoder rates (1,967 in *Expert* and 894 in *Middle*) were equivalent to the top 6.5% and 34.1% positions among 7,671 registered players, respectively. Seven additional subjects were scanned but not included in the analysis because one showed neurological abnormality in MRI images, three retired the experiment without full completion, three showed strongly-biased behavioral responses judged when the behavioral performance of one or more *category* did not

reach chance-level in the training experiments, signaling the strong response bias sticking to a specific choice. This study was approved by the Ethics Committees of NICT and NAIST and subjects gave written informed consent for participation.

Source code selection and normalization.

Source code snippets were collected from an open codeset provided by AIZU ONLINE JUDGE (http://judge.u-aizu.ac.jp/onlinejudge/), an online judge system where lots of programming problems are listed and everyone can submit their own source code to answer those online. We selected four functional categories (*category*) and eleven subordinate concrete algorithms (*subcategory*) based on two popular textbooks about computer algorithms[65,66] (see Fig.1a). We then searched in the open codeset for Java code snippets implementing one of the selected algorithms and found 1251 candidates. The reasons why we focused on Java in this study were because the language has been one of the most famous programming languages[67] and prior fMRI studies on programmers also used Java code snippets as experimental stimuli[17–19]. To meet the screen size constraint in the MRI scanner, we excluded code snippets with lines of code (LOC) more than 30 and characters per line (CPL) more than 120. From all remaining snippets, we created a set of 72 code snippets with minimum deviations of LOC and CPL to minimize visual variation as experimental stimuli; the mean and standard deviation of LOC and CPL were 26.4 ± 2.4 and 59.3 ± 17.1, respectively. In the codeset, 18 snippets each belonged to one of the *category* classes and six snippets each belonged to one of the *subcategory* classes except for "linear search" class with twelve snippets (see Supplementary Table.3 for detailed statics of the codeset). The indentation styles of all code snippets were normalized by replacing a tab-space with two white-spaces and all user-defined functions were renamed to neutral like "function1" because some of them indicated their algorithms explicitly (see Supplementary Figure 1 for example snippets used in the experiment). Finally, we verified all code snippets had no syntax error and run correctly without run-time error.

Experimental design.

The fMRI experiment consisted of six separate runs (9 min 52 sec for each run). Each run contained 36 trials of the program categorization task (Fig.1b) plus one dummy trial to avoid undesirable effects of MRI signal instability. We used 72 types of Java code snippets as stimuli and each snippet was presented three times in total through the whole experiment, but the same snippet appeared only once in a run. We employed PsychoPy[68] (version 1.85.1) to display the code snippets in white text and gray background without any syntax highlighting to minimize visual variations. In each trial of program categorization tasks, a Java code snippet was displayed for ten seconds after a fixation-cross presentation for two seconds. Subjects then classified the given code snippet into one of four *category* classes within four seconds by pressing a button placed under the right hand. To clarify classification criteria, a brief explanation about each *category* class was provided before the experiment started (see supplementary information for the detailed descriptions). The presentation order of the code snippets was randomized under balancing the number of exemplars for each *category* class across runs. The corresponding buttons for each answer choice were also randomized across trials to avoid linking a specific answer choice with a specific finger movement. Subjects were allowed to take a break between runs and to quit the fMRI experiment at any time.

To mitigate potential noises caused by task unfamiliarity, every subject conducted a training experiment within ten days before the fMRI experiment. The training experiment consisted of three separate runs with the same settings as the fMRI experiment. A different set of 72 Java code snippets implementing the same algorithms was used as stimuli; each snippet was presented one or two times but the same snippet appeared only once in a run. In addition, all subjects took a post-MRI experiment within ten days after the fMRI experiments for assessment of individual ability to *subcategory* categorizations. The post-MRI experiment consisted of two separate runs using the same codeset as the fMRI experiment. Before the post-MRI experiments started, we revealed the existence of *subcategory* and provided brief descriptions about each *subcategory* class (see supplementary information for the detailed descriptions). Subjects classified the given code snippet from two or three choices of *subcategory* classes according to its superordinate *category*, e.g. 'bubble sort', 'insertion sort', 'selection sort' were displayed when the snippet in 'sort' category was presented. The training and post-MRI experiments were performed outside of the MRI scanner. For all experiments, we calculated behavioral performance as a ratio of correct-answer-trials in all-trials; unanswered trials were regarded as 'incorrect' for this calculation. Chance-level behavioral performance was 25% in the training and fMRI experiments and 37.25% in the post-MRI experiment adjusted for imbalanced numbers of answer choices.

MRI data acquisition.

fMRI data were collected using 3-Tesla Siemens MAGNETOM Prisma scanner with a 64-channel head coil located at CiNet. T2*-weighted multiband gradient echo-EPI sequences were performed to acquire functional images covering the entire brain (repetition time (TR) = 2000 ms, echo time (TE) = 30 ms, flip angle = 75°, field of view (FOV) = 192 × 192 mm, slice thickness = 2 mm, slice gap = 0 mm, voxel size = $2 \times 2 \times 2.01$ mm, multi-band factor = 3). A T1-weighted magnetization-prepared rapid acquisition with gradient-echo sequence was also performed to acquire fine-structural images of the entire head (TR = 2530 ms, TE = 3.26 ms, flip angle = 9°, FOV = 256 × 256 mm, slice thickness = 1 mm, slice gap = 0 mm, voxel size = $1 \times 1 \times 1$ mm).

MRI data preprocessing.

We used the Statistical Parameter Mapping toolbox (SPM12, http://www.fil.ion.ucl.ac.uk/spm/) for preprocessing. The first eight scans in dummy trials for each run were discarded to avoid MRI signal instability. The functional scans were aligned to the first volume in the fourth run to remove movement artifacts. They were then slice-time corrected and co-registered to the whole-head T1 structural image. Both anatomical and functional images were spatially normalized into the standard Montreal Neurological 152-brain average template space and resampled to a voxel size of $2 \times 2 \times 2$ mm. MRI signals at each voxel were high-pass–filtered with a cutoff period of 128 seconds to remove low-frequency drifts. A thick gray matter mask was obtained from the normalized anatomical images of all subjects to select the voxels within neuronal tissue using the SPM Masking Toolbox[69]. For each subject independently, we then fitted a general linear model (GLM) to estimate voxel-level parameters (β) linking recorded MRI signals and conditions of source code presentations in each trial. The fixation and response phases in each trial were not explicitly modeled. The model also included motion realignment parameters to regress-out signal variations due to head motion. Finally, 216 beta estimate maps (36 trials $\times$ 6 runs) per subject were yielded and used as input for the following multivariate pattern analysis.

Decoding functional category of source code.

We used a whole-brain searchlight analysis[27] to examine where significant decoding accuracies exists in a data-driven manner. A four-voxel-radius sphered searchlight, covering 251 voxels at once, was systematically shifted throughout the brain and decoding accuracy was quantified on each searchlight location. We employed The Decoding Toolbox[70] (version 3.99) and a linear-kernel SVM classifier as implemented in LIBSVM[71] (version 3.17) to decode the functional *category* and *subcategory* of seen source code from fMRI activity. The SVM classifier was trained and evaluated using a leave-one-run-out cross-validation procedure, which iteratively treated data in a single run for test and others for training. In each fold, training data was first scaled to zero-mean and unit variance by z-transform and test data was scaled using the estimated scaling parameters. We then applied outlier reduction using [-3, +3] as cut-off values and all scaled signals larger than the upper cut-off or smaller than the lower cut-off were set to the closest value of these limits. After that, the SVM classifier was trained with three cost parameter candidates [0.1, 1, 10] and the best parameter was automatically chosen by grid search in nested cross-validations. Due to the constraint of the high computational load of searchlight analysis, we here adopted the relatively small set of cost parameters. Finally, the trained classifier predicted *category* or *subcategory* of seen source code from the leave-out test data and decoding accuracy was calculated as a ratio of correct-classifications out of all-classifications. Note that corrected misclassification cost weights were used in *subcategory* decoding to compensate for the imbalanced number of exemplars across *subcategory* classes.

The training and evaluation procedure was performed for each subject independently and we obtained a decoding accuracy map per subject as a result. We then conducted second-level analyses to examine the significance of decoding accuracies and correlations between individual decoding and behavioral performances. For this purpose, the resulted decoding accuracy maps were spatially smoothed using a Gaussian kernel of 6 mm full-width at half maximum (FWHM) and submitted to random effects analysis. We again employed SPM12 for the random effects analysis and tested the significance of group-level decoding accuracy and Pearson's correlation coefficient between individual decoding accuracies and behavioral performances. We adopted a relatively strict statistical threshold of voxel-level $p < 0.05$ FWE-corrected for decoding accuracy tests and a basic threshold of voxel-level $p < 0.001$ uncorrected and cluster-level $p < 0.05$ FWE-corrected for correlation tests. For null hypotheses, chance-level decoding accuracies, i.e. 25% in *category* decoding and 9.72% in *subcategory* decoding adjusted for imbalanced numbers of exemplar, and zero correlation were used. Our experimental procedure followed the expert performance approach[21,72], in which experts' superior performance was first captured by a representative laboratory task and then neuroimaging data was examined to understand how the brain accommodates such behavioral superiority.

Data and code availability.

The experimental data and code used in the present study are available from our repository: https://github.com/Yoshiharu-Ikutani/DecodingCodeFromTheBrain.

References

1. Li, P. L., Ko, A. J. & Zhu, J. What makes a great software engineer? In *Proceedings of the 37th International Conference on Software Engineering-Volume 1*, 700–710 (IEEE Press, 2015).

2. Baltes, S. & Diehl, S. Towards a theory of software development expertise. In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2018, 187–200 (ACM, 2018).

3. Heintz, F., Mannila, L. & Färnqvist, T. A review of models for introducing computational thinking, computer science and computing in k-12 education. In *2016 IEEE Frontiers in Education conference (FIE)*, 1–9 (IEEE, 2016).

4. Papavlasopoulou, S., Sharma, K. & Giannakos, M. N. How do you feel about learning to code? investigating the effect of children's attitudes towards coding using eye-tracking. *Int. J. Child-Computer Interact.* **17**, 50–60 (2018).
5. Sackman, H., Erikson, W. J. & Grant, E. E. Exploratory experimental studies comparing online and offline programing performance. Tech. Rep., SYSTEM DEVELOPMENT CORP SANTA MONICA CA (1966).
6. Boehm, B. W. Understanding and controlling software costs. *J. Parametr.* **8**, 32–68 (1988).
7. DeMarco, T. & Lister, T. *Peopleware: Productive Projects and Teams (3rd Edition)* (Addison-Wesley Professional, 2013), 3rd edn.
8. Vessey, I. Expertise in debugging computer programs: A process analysis. *Int. J. Man-Machine Stud.* **23**, 459–494 (1985).
9. Koenemann, J. & Robertson, S. P. Expert problem solving strategies for program comprehension. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 125–130 (ACM, 1991).
10. Guindon, R. Knowledge exploited by experts during software system design. *Int. J. Man-Machine Stud.* **33**, 279–304, DOI: 10.1016/S0020-7373(05)80120-8 (1990).
11. Fix, V., Wiedenbeck, S. & Scholtz, J. Mental representations of programs by novices and experts. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, 74–79 (ACM, 1993).
12. Von Mayrhauser, A. & Vans, A. M. Program comprehension during software maintenance and evolution. *Computer* 44–55 (1995).
13. Crosby, M. E., Scholtz, J. & Wiedenbeck, S. The roles beacons play in comprehension for novice and expert programmers. In *14th Workshop of the Psychology of Programming Interest Group*, 58–73 (2002).
14. Uwano, H., Nakamura, M., Monden, A. & Matsumoto, K.-i. Analyzing individual performance of source code review using reviewers' eye movement. In *Proceedings of the 2006 symposium on Eye tracking research & applications*, 133–140 (ACM, 2006).
15. Sharif, B., Falcone, M. & Maletic, J. I. An eye-tracking study on the role of scan time in finding source code defects. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, 381–384 (ACM, 2012).
16. Busjahn, T. *et al.* Eye movements in code reading: Relaxing the linear order. In *2015 IEEE 23rd International Conference on Program Comprehension*, 255–265 (IEEE, 2015).
17. Siegmund, J. *et al.* Understanding understanding source code with functional magnetic resonance imaging. In *Proceedings of the 36th International Conference on Software Engineering*, 378–389 (ACM, 2014).
18. Siegmund, J. *et al.* Measuring neural efficiency of program comprehension. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, 140–150 (ACM, 2017).
19. Peitek, N. *et al.* A look into programmers' heads. *IEEE Transactions on Softw. Eng.* (2018).
20. Floyd, B., Santander, T. & Weimer, W. Decoding the representation of code in the brain: An fmri study of code review and expertise. In *2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE)*, 175–186, DOI: 10.1109/ICSE.2017.24 (2017).
21. Bilalić, M. *The Neuroscience of Expertise*. Cambridge Fundamentals of Neuroscience in Psychology (Cambridge University Press, 2017).
22. Wasik, S., Antczak, M., Badura, J., Laskowski, A. & Sternal, T. A survey on online judge systems and their applications. *ACM Comput. Surv. (CSUR)* **51**, 3 (2018).
23. Bilalić, M., Grottenthaler, T., Nägele, T. & Lindig, T. The faces in radiological images: fusiform face area supports radiological expertise. *Cereb. Cortex* **26**, 1004–1014 (2016).
24. de Borst, A. W., Valente, G., Jääskeläinen, I. P. & Tikka, P. Brain-based decoding of mentally imagined film clips and sounds reveals experience-based information patterns in film professionals. *Neuroimage* **129**, 428–438 (2016).
25. Martens, F., Bulthé, J., van Vliet, C. & de Beeck, H. O. Domain-general and domain-specific neural changes underlying visual expertise. *NeuroImage* **169**, 80–93 (2018).
26. Gomez, J., Barnett, M. & Grill-Spector, K. Extensive childhood experience with pokémon suggests eccentricity drives organization of visual cortex. *Nat. human behaviour* **3**, 611 (2019).
27. Kriegeskorte, N., Goebel, R. & Bandettini, P. Information-based functional brain mapping. *Proc. Natl. Acad. Sci.* **103**, 3863–3868 (2006).

28. Etzel, J. A., Zacks, J. M. & Braver, T. S. Searchlight analysis: promise, pitfalls, and potential. *Neuroimage* **78**, 261–269 (2013).

29. Tanaka, J. W. & Taylor, M. Object categories and expertise: Is the basic level in the eye of the beholder? *Cogn. psychology* **23**, 457–482 (1991).

30. Castelhano, J. *et al.* The role of the insula in intuitive expert bug detection in computer code: an fmri study. *Brain imaging behavior* 1–15 (2018).

31. Peitek, N. *et al.* Simultaneous measurement of program comprehension with fmri and eye tracking: A case study. In *Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, 24:1–24:10 (ACM, 2018).

32. Demonet, J.-F. *et al.* The anatomy of phonological and semantic processing in normal subjects. *Brain* **115**, 1753–1768 (1992).

33. Thompson-Schill, S. L., D'Esposito, M., Aguirre, G. K. & Farah, M. J. Role of left inferior prefrontal cortex in retrieval of semantic knowledge: a reevaluation. *Proc. Natl. Acad. Sci.* **94**, 14792–14797 (1997).

34. Simmons, A., Miller, D., Feinstein, J. S., Goldberg, T. E. & Paulus, M. P. Left inferior prefrontal cortex activation during a semantic decision-making task predicts the degree of semantic organization. *Neuroimage* **28**, 30–38 (2005).

35. Price, C. J. A review and synthesis of the first 20 years of pet and fmri studies of heard speech, spoken language and reading. *Neuroimage* **62**, 816–847 (2012).

36. Rodd, J. M., Davis, M. H. & Johnsrude, I. S. The neural mechanisms of speech comprehension: fmri studies of semantic ambiguity. *Cereb. Cortex* **15**, 1261–1269 (2005).

37. Kuhl, B. A., Dudukovic, N. M., Kahn, I. & Wagner, A. D. Decreased demands on cognitive control reveal the neural processing benefits of forgetting. *Nat. neuroscience* **10**, 908 (2007).

38. Whitney, C., Jefferies, E. & Kircher, T. Heterogeneity of the left temporal lobe in semantic representation and control: priming multiple versus single meanings of ambiguous words. *Cereb. cortex* **21**, 831–844 (2010).

39. Bookheimer, S. Y., Zeffiro, T. A., Blaxton, T., Gaillard, W. & Theodore, W. Regional cerebral blood flow during object naming and word reading. *Hum. Brain Mapp.* **3**, 93–106 (1995).

40. Philipose, L. E. *et al.* Neural regions essential for reading and spelling of words and pseudowords. *Annals Neurol. Off. J. Am. Neurol. Assoc. Child Neurol. Soc.* **62**, 481–492 (2007).

41. Stoeckel, C., Gough, P. M., Watkins, K. E. & Devlin, J. T. Supramarginal gyrus involvement in visual word recognition. *Cortex* **45**, 1091–1096 (2009).

42. Wagner, A. D., Shannon, B. J., Kahn, I. & Buckner, R. L. Parietal lobe contributions to episodic memory retrieval. *Trends cognitive sciences* **9**, 445–453 (2005).

43. Vilberg, K. L. & Rugg, M. D. Memory retrieval and the parietal cortex: a review of evidence from a dual-process perspective. *Neuropsychologia* **46**, 1787–1799 (2008).

44. O'Connor, A. R., Han, S. & Dobbins, I. G. The inferior parietal lobule and recognition memory: expectancy violation or successful retrieval? *J. Neurosci.* **30**, 2924–2934 (2010).

45. Chelazzi, L., Miller, E. K., Duncan, J. & Desimone, R. A neural basis for visual search in inferior temporal cortex. *Nature* **363**, 345 (1993).

46. Nobre, A. C., Allison, T., McCarthy, G. *et al.* Word recognition in the human inferior temporal lobe. *Nature* **372**, 260–263 (1994).

47. Kriegeskorte, N. *et al.* Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* **60**, 1126–1141 (2008).

48. Corbetta, M., Patel, G. & Shulman, G. L. The reorienting system of the human brain: from environment to theory of mind. *Neuron* **58**, 306–324 (2008).

49. Japee, S., Holiday, K., Satyshur, M. D., Mukai, I. & Ungerleider, L. G. A role of right middle frontal gyrus in reorienting of attention: a case study. *Front. systems neuroscience* **9**, 23 (2015).

50. Bilalić, M., Kiesel, A., Pohl, C., Erb, M. & Grodd, W. It takes two–skilled recognition of objects engages lateral areas in both hemispheres. *PLoS One* **6**, e16202 (2011).

51. Tanaka, S., Michimata, C., Kaminaga, T., Honda, M. & Sadato, N. Superior digit memory of abacus experts: an event-related functional mri study. *Neuroreport* **13**, 2187–2191 (2002).

52. Hanakawa, T., Honda, M., Okada, T., Fukuyama, H. & Shibasaki, H. Neural correlates underlying mental calculation in abacus experts: a functional magnetic resonance imaging study. *Neuroimage* **19**, 296–307 (2003).

53. Tanaka, S. *et al.* Abacus in the brain: a longitudinal functional mri study of a skilled abacus user with a right hemispheric lesion. *Front. psychology* **3**, 315 (2012).

54. Moore, C. J. & Price, C. J. Three distinct ventral occipitotemporal regions for reading and object naming. *Neuroimage* **10**, 181–192 (1999).

55. Burton, M. W., Noll, D. C. & Small, S. L. The anatomy of auditory word processing: individual variability. *Brain language* **77**, 119–131 (2001).

56. Diedrichsen, J. & Kriegeskorte, N. Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS computational biology* **13**, e1005508 (2017).

57. Horikawa, T. & Kamitani, Y. Generic decoding of seen and imagined objects using hierarchical visual features. *Nat. communications* **8**, 15037 (2017).

58. Nishida, S. & Nishimoto, S. Decoding naturalistic experiences from human brain activity via distributed representations of words. *Neuroimage* **180**, 232–242 (2018).

59. Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E. & Gallant, J. L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **532**, 453 (2016).

60. Pereira, F. *et al.* Toward a universal decoder of linguistic meaning from brain activation. *Nat. communications* **9**, 963 (2018).

61. Alon, U., Levy, O. & Yahav, E. code2seq: Generating sequences from structured representations of code. *arXiv preprint arXiv:1808.01400* (2018).

62. Zhang, J. *et al.* A novel neural source code representation based on abstract syntax tree. In *2019 IEEE/ACM 41th International Conference on Software Engineering (ICSE)*, 783–794 (2019).

63. Alon, U., Zilberstein, M., Levy, O. & Yahav, E. code2vec: Learning distributed representations of code. *Proc. ACM on Program. Lang.* **3**, 40 (2019).

64. Oldfield, R. C. The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia* **9**, 97–113 (1971).

65. Cormen, T. H., Leiserson, C. E., Rivest, R. L. & Stein, C. *Introduction to Algorithms, Third Edition* (The MIT Press, 2009), 3rd edn.

66. Sedgewick, R. & Wayne, K. *Algorithms* (Addison-Wesley Professional, 2011), 4th edn.

67. Elliott, T. The state of the octoverse: top programming languages of 2018. https://github.blog/2018-11-15-state-of-the-octoverse-top-programming-languages/ (2018). Opened: 2018-11-15, Accessed: 2019-09-01.

68. Peirce, J. W. Psychopy—psychophysics software in python. *J. neuroscience methods* **162**, 8–13 (2007).

69. Ridgway, G. R. *et al.* Issues with threshold masking in voxel-based morphometry of atrophied brains. *Neuroimage* **44**, 99–111 (2009).

70. Hebart, M. N., Görgen, K. & Haynes, J.-D. The decoding toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. *Front. neuroinformatics* **8**, 88 (2015).

71. Chang, C.-C. & Lin, C.-J. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems technology (TIST)* **2**, 27 (2011).

72. Ericsson, K. A. & Smith, J. *Toward a general theory of expertise: Prospects and limits* (Cambridge University Press, 1991).

73. Xia, M., Wang, J. & He, Y. Brainnet viewer: a network visualization tool for human brain connectomics. *PloS one* **8**, e68910 (2013).

Acknowledgements

We thank Takao Nakagawa and Hidetake Uwano for helpful comments on the initial study design. This work was supported by JSPS KAKENHI Grant Number JP15H05311, JP16H05857, JP16H06569, JP17H01797, JP18K18108, JP18K18141, JP18J22957, and JST ERATO Grant Number JPMJER1801.

Author contributions

I.Y., T.K., S.Nishida, H.H., and S.Nishimoto designed the study. I.Y., T.K., and H.H. recruited subjects. I.Y., T.K., S.Nishida, and S.Nishimoto conducted experiments. I.Y., T.K., and S.Nishida analyzed data. All authors contributed interpreting results. I.Y., T.K., and S.Nishida wrote the manuscript with input from other co-authors.

Competing interest

The named authors declare no competing financial interests.

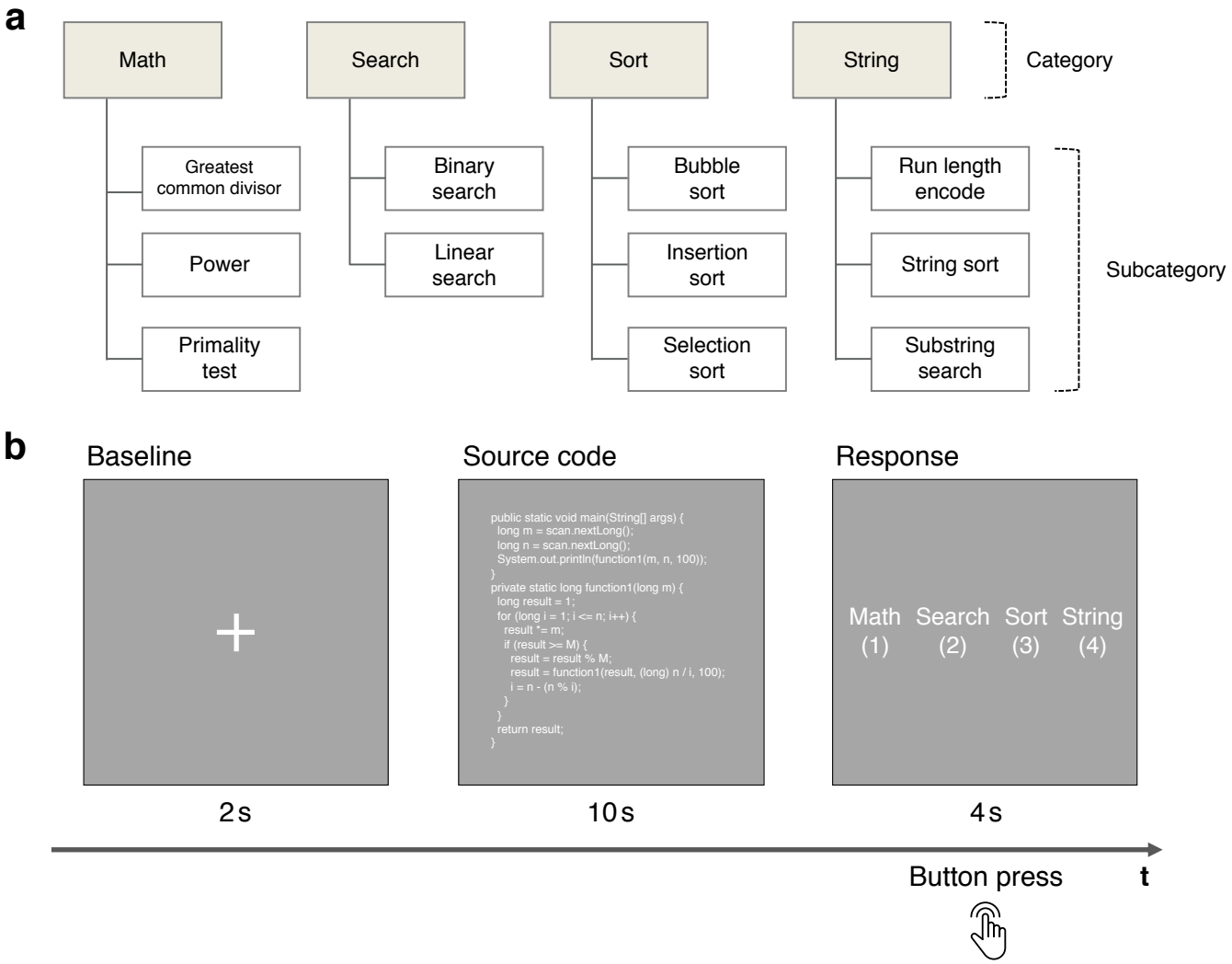

Figure 1. Experimental design. (**a**) functional structure of source code used in this study. Category and Subcategory layers each represented abstract functionality and concrete algorithms based on two popular textbooks of programming. Every code snippet used in this study belonged to one subcategory class and its corresponding category class. (**b**) Program categorization task. After a fixation-cross presentation for two seconds, a Java code snippet was displayed for ten seconds in white text without any syntax highlight. Then, subjects responded the category of given code snippet by pressing a button.

Table 1 : Demographic information of recruited subjects

Recruiting criteria	N	Gender (M/F)	Age	AtCoder rate	Programming experience	Java experience	Competitive programming experience
Expert	10	10 / 0	22.6 ± 1.1	1969 ± 467	6.9 ± 2.8	2.8 ± 2.4	4.1 ± 2.6 *
Middle	10	9 / 1	22.5 ± 0.8	894 ± 175	4.8 ± 1.7 *	1.1 ± 0.8	1.3 ± 0.8
Novice	10	9 / 1	21.7 ± 1.2	NA	2.8 ± 0.6 *	1.4 ± 1.0	NA

Numerics from 4th to last columns denote 'MEAN ± SD'; Single asterisk indicates $p < 0.05$ FDR corrected, using two-sample *t*-test.

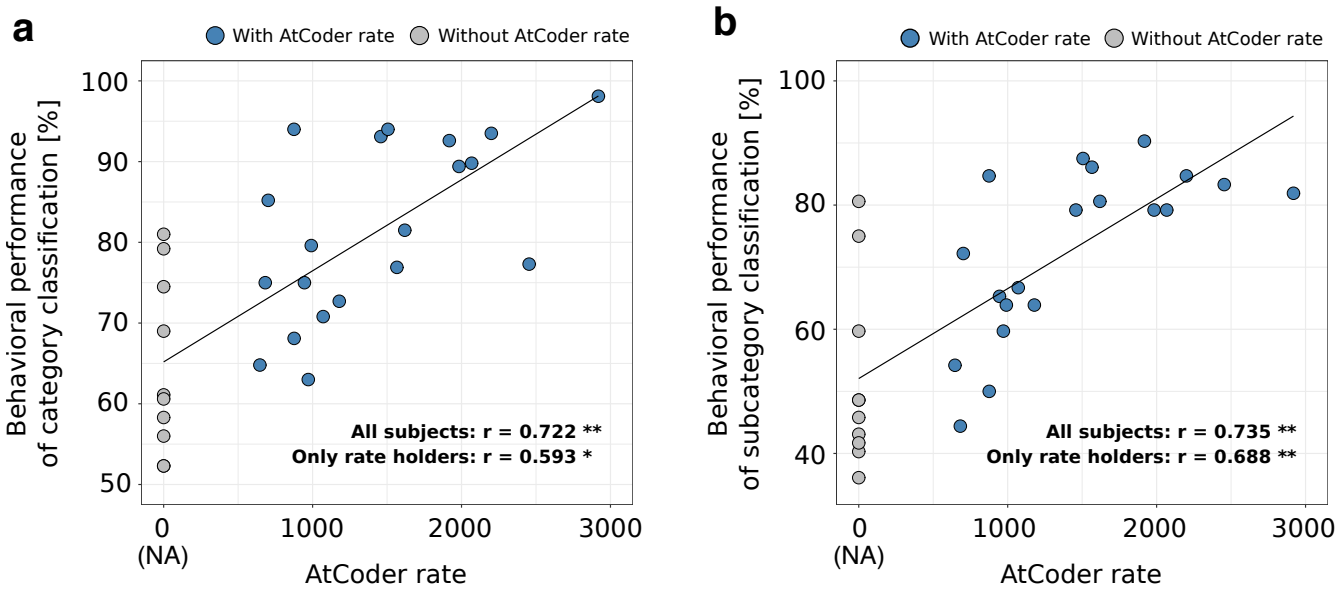

Figure 2. Behavioral performances of the program categorization task significantly correlated with the indicator of programming expertise in competitive programming contests. (**a**) Scatter plot of behavioral performances of category classifications against the expertise indicator (i.e. AtCoder rate). (**b**) Scatter plot of behavioral performances of subcategory classifications against the expertise indicator. Each dot represents an individual subject. One-sample t-tests were used to check significance of the correlation coefficients (r) between the expertise indicator and behavioral performances; *, $p < 0.05$ and **, $p < 0.005$. The solid lines indicates a fitted regression line estimated from all subject data.

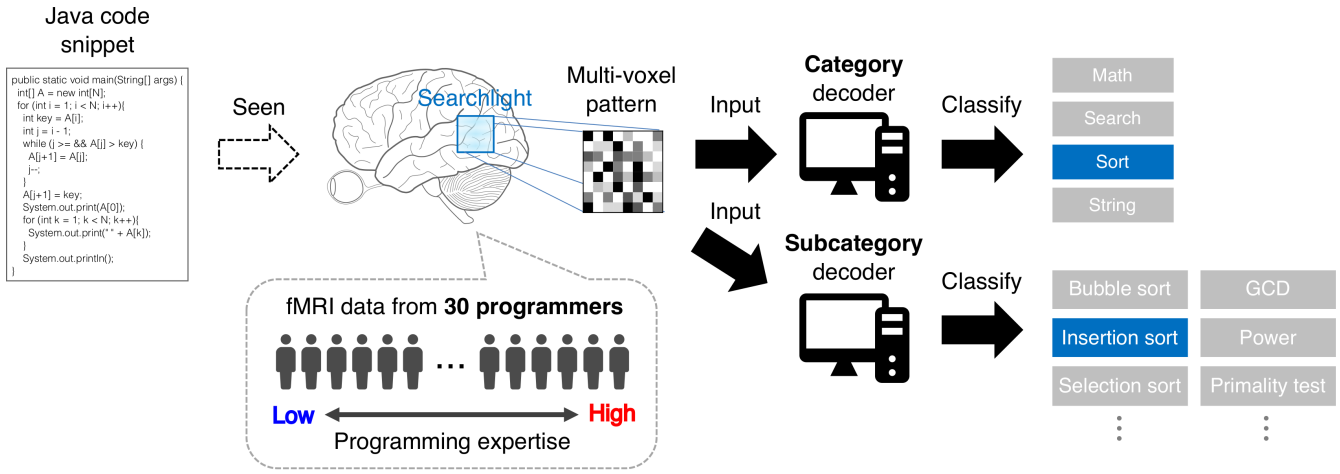

Figure 3. Decoding functional category of source code from the brain. Overview of the decoding framework. Functional MRI data was collected from 30 subjects with different levels of programming expertise while they performed the program categorization task. We employed whole-brain searchlight analysis[27] to explore the potential loci of programming expertise. For each searchlight location, a linear-kernel SVM classifier (decoder) was trained on multi-voxel patterns to classify *category* or *subcategory* of given Java code snippets. These procedures were performed independently for each subject. Decoding accuracies were calculated as a ratio of correct-classifications out of all-classifications.

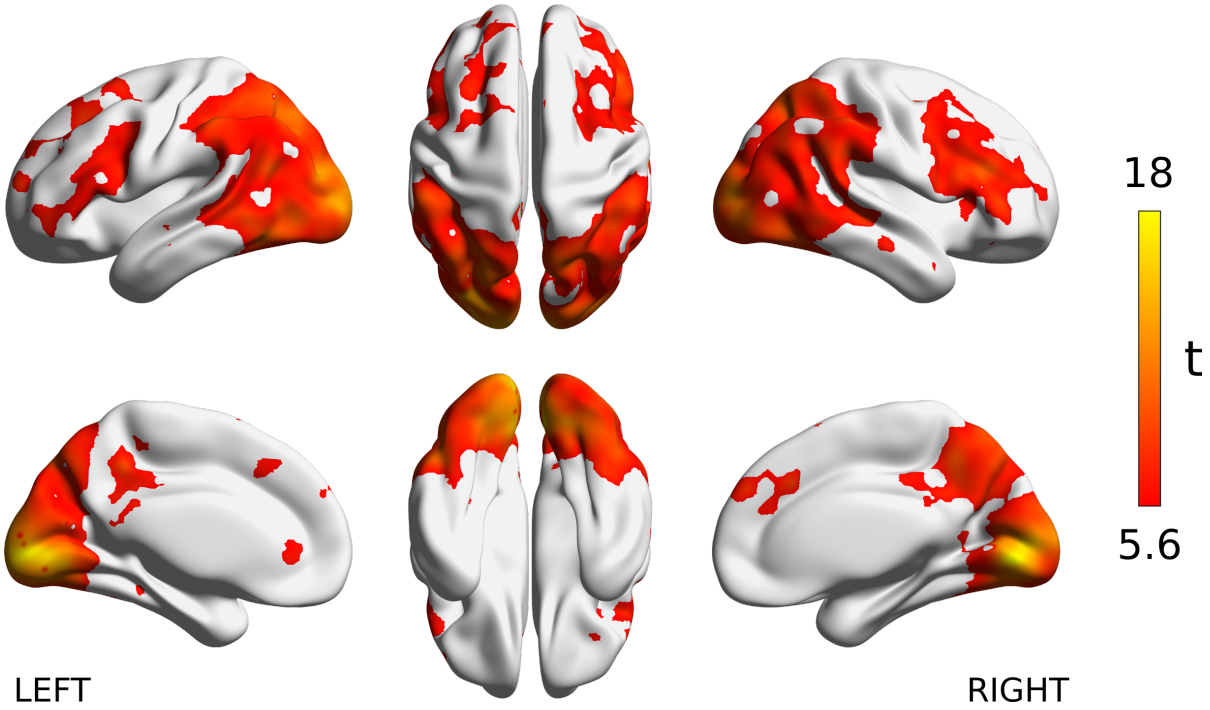

Figure 4. Decoding accuracy for functional category of source code. Significant searchlight locations estimated from all subject data (N = 30). Heat colored voxels denote the centers of searchlights with significant decoding accuracy (voxel-level $p < 0.05$, FWE corrected). The brain surface visualizations were performed using BrainNet viewer, version 1.61[73].

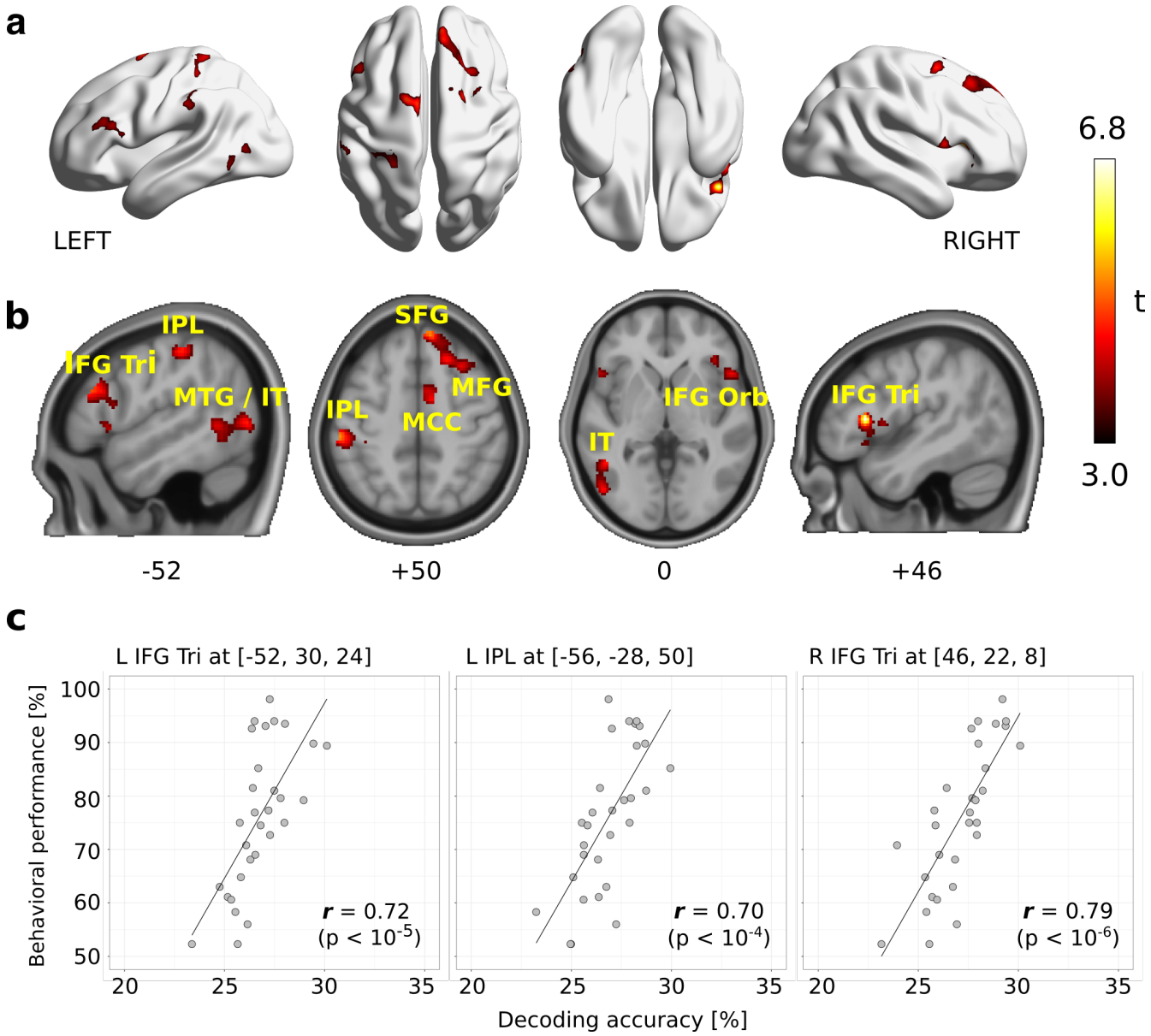

Figure 5. Searchlight-based correlation analysis between behavioral performances and decoding accuracies. (**a**) Locations of searchlight showing significant correlations. (**b**) Slice-wise visualizations of the significant clusters using bspmview (http://www.bobspunt.com/software/bspmview). Significance was determined by a threshold of voxel-level $p < 0.001$ and cluster-level $p < 0.05$, FWE corrected for the whole brain. (**c**) Scatter plots of peak correlations between decoding accuracies and behavioral performances. Each dot represents an individual subject data. Correlation coefficients (r) and uncorrected p values are shown in bottom-right of each plot. See Supplementary Table.4 and Supplementary Fig.2 for all significant clusters and peak correlations. Abbreviations: IFG Tri, Inferior frontal gyrus pars triangularis; IFG Orb, Inferior frontal gyrus pars orbitalis; SFG, Superior frontal gyrus; MFG, middle frontal gyrus; IPL, Inferior parietal lobule; MTG, Middle temporal gyrus; IT, Inferior temporal gyrus; MCC, medial cingulate cortex.

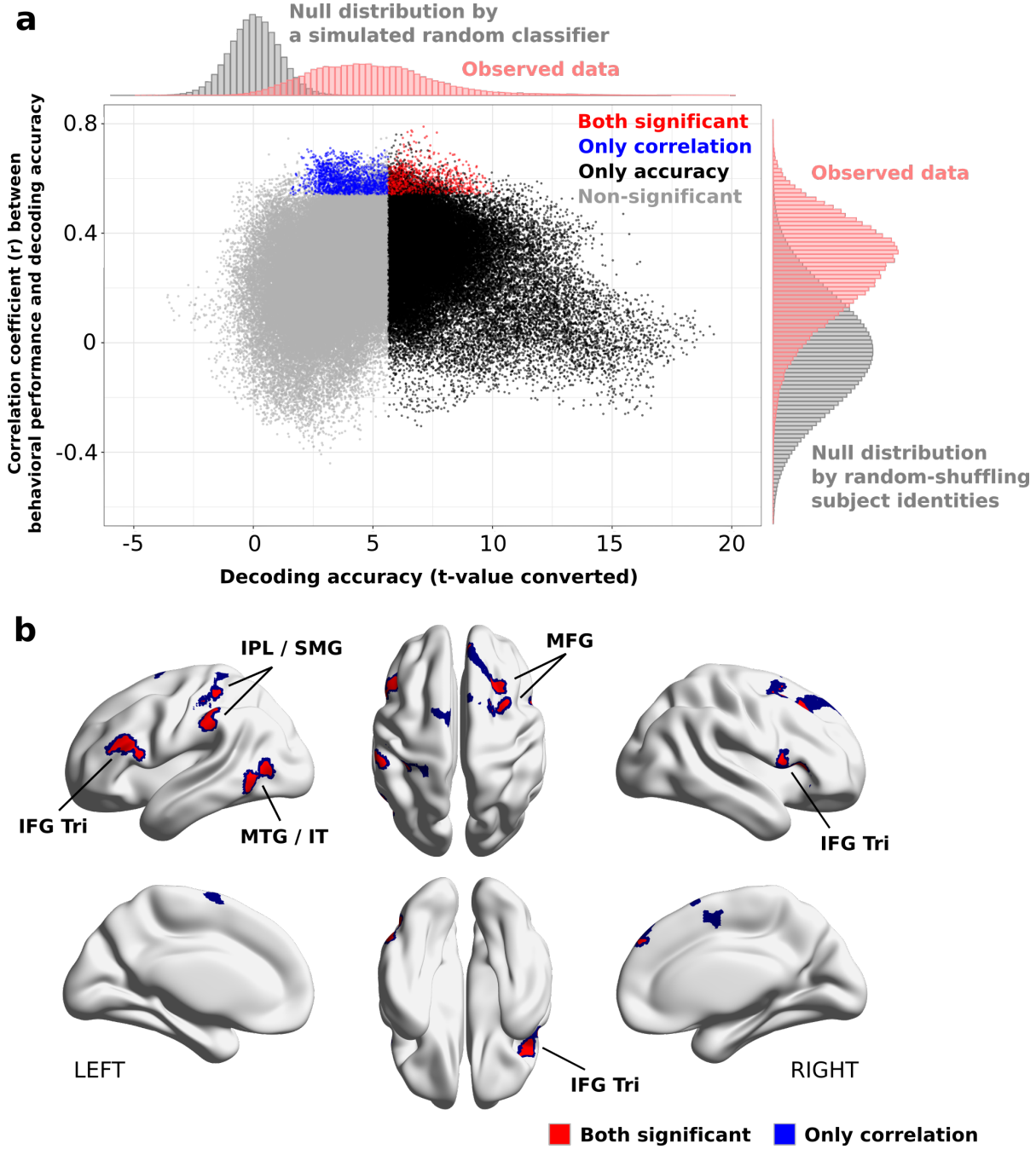

Figure 6. Identifying searchlight centers that showed both significant decoding accuracy and significant correlation to individual behavioral performances. (**a**) Scatter plot of searchlight results. X-axis shows t-values calculated from all subjects' decoding accuracies on each searchlight locations. Y-axis indicates correlation coefficients between decoding accuracies and behavioral performances. Red-colored dots denote the searchlights showing both significant decoding accuracy and correlation, while blue and black denote those only showed significant decoding accuracy or correlations. Non-significant searchlights were colored in gray. The observed distributions of decoding accuracies and correlations are respectively shown on top- and right-sides of the figure accompanied with null distributions calculated by randomized simulations. (**b**) Locations of searchlight centers that showed both significant decoding accuracy and significant correlations to individual behavioral performances. Abbreviations: SMG, Supramarginal gyrus; others are same as used in Fig.5.

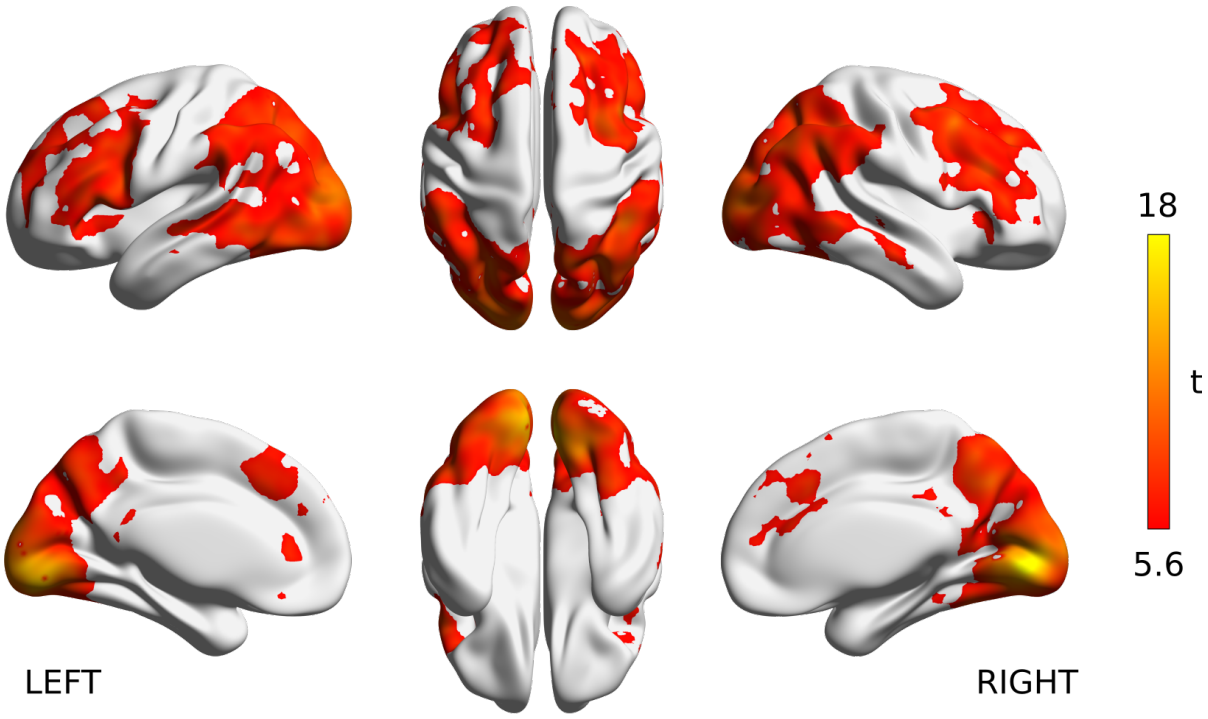

Figure 7. Decoding accuracy for subcategory of source code. Searchlight locations showing significant subcategory decoding accuracy than chance estimated from all subject data (N = 30). Heat colored voxels denote the centers of searchlights with significant subcategory decoding accuracy (voxel-level $p < 0.05$, FWE corrected).

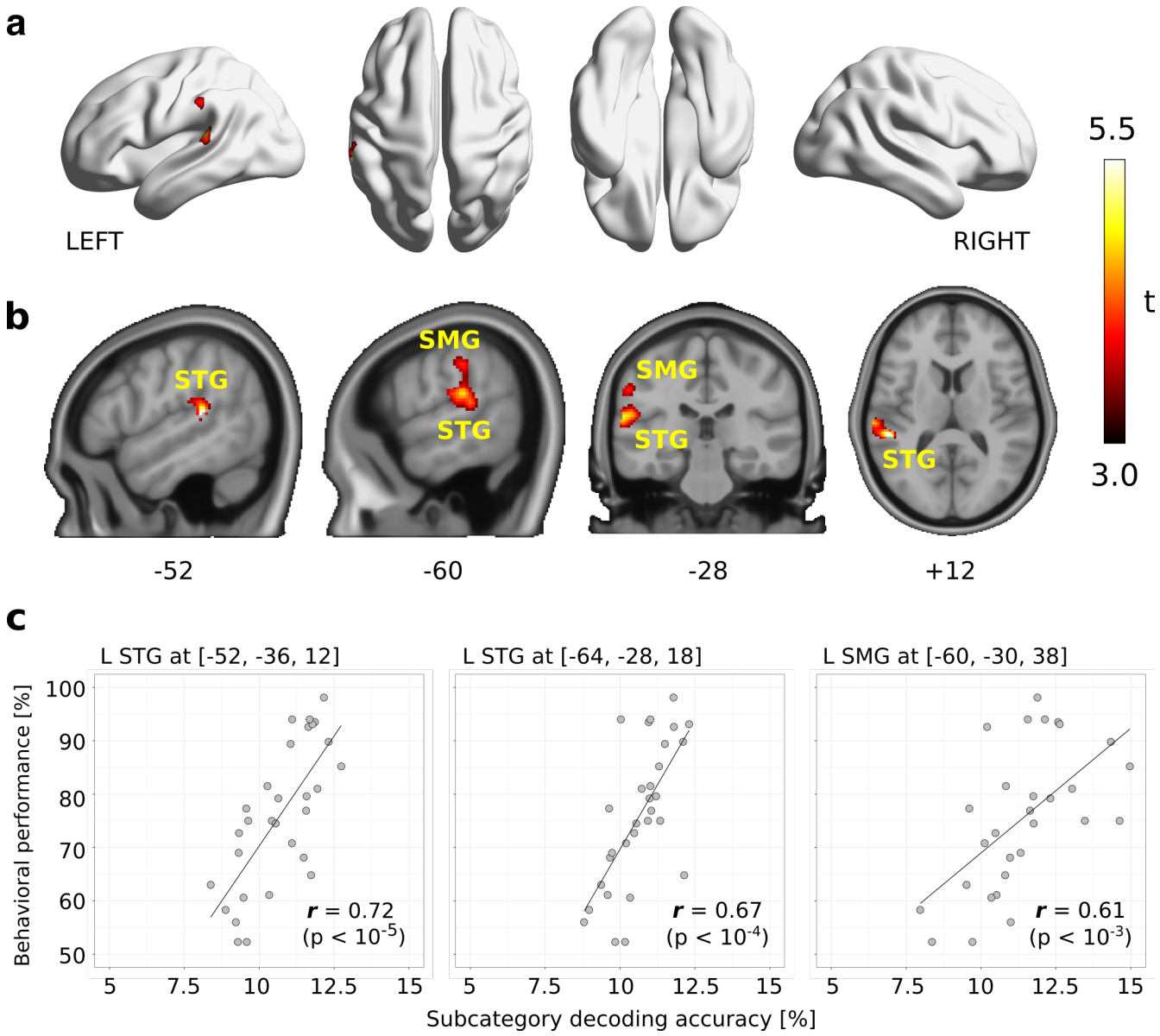

Figure 8. Searchlight-based correlation analysis between behavioral performances and subcategory decoding accuracies. (**a**) Locations of searchlight showing significant correlations. (**b**) Slice-wise visualizations of the significant clusters. Significance was determined by a threshold of voxel-level $p < 0.001$ and cluster-level $p < 0.05$, FWE corrected for the whole brain. (**c**) Scatter plots of peak correlations between decoding accuracies and behavioral performances. Each dot represents an individual subject data. Correlation coefficients (r) and uncorrected p values are shown in bottom-right of each plot. Only one cluster (extent = 501 voxels) had significant correlation in this analysis and three peak correlations in the cluster were shown here. Abbreviations: STG, Superior temporal gyrus; SMG, Supramarginal gyrus.

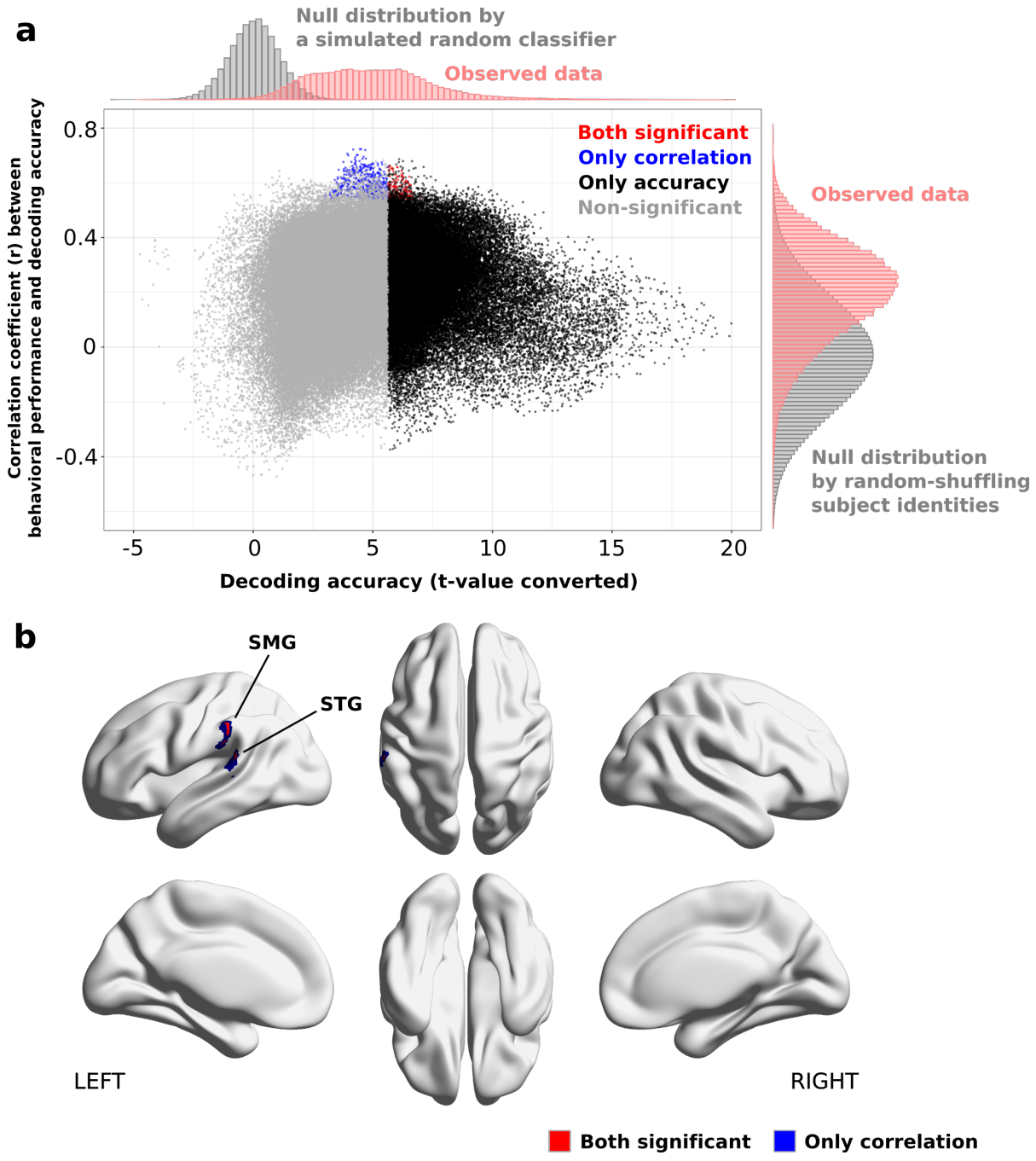

Figure 9. Identifying searchlight centers that showed both significant subcategory decoding accuracy and significant correlation to individual behavioral performances. (**a**) Scatter plot of searchlight results. X-axis shows t-values calculated from all subjects' decoding accuracies on each searchlight locations. Y-axis indicates correlation coefficients between subcategory decoding accuracies and behavioral performances. Red-colored dots denote the searchlights showing both significant decoding accuracy and correlation, while blue and black denote those only showed significant decoding accuracy or correlations. Non-significant searchlights were colored in gray. The observed distributions of decoding accuracies and correlations are respectively shown on top- and right-sides of the figure accompanied with null distributions calculated by randomized simulations. (**b**) Locations of searchlight centers that showed both significant subcategory decoding accuracy and significant correlations to individual behavioral performances. Abbreviations are same as used in Fig.8.