eNeuro

Research Article: New Research | Cognition and Behavior

Expert programmers have fine-tuned cortical representations of source code

https://doi.org/10.1523/ENEURO.0405-20.2020

Cite as: eNeuro 2020; 10.1523/ENEURO.0405-20.2020

Received: 18 September 2020
Revised: 14 November 2020
Accepted: 1 December 2020

Manuscript Title Page

1. Manuscript Title (50 word maximum)

Expert programmers have fine-tuned cortical representations of source code

2. Abbreviated Title (50 character maximum)

Programming expertise and programmers' brain

3. List of all Author Names and Affiliations

Yoshiharu Ikutani [a], Takatomi Kubo [a], Satoshi Nishida [b], Hideaki Hata [a], Kenichi Matsumoto [a], Kazushi Ikeda [a], Shinji Nishimoto [b]

[a] Graduate School of Science and Technology, Division of Information Science, Nara Institute of Science and Technology (NAIST), Ikoma, Nara 630-0192, Japan

[b] Center for Information and Neural Networks (CiNet), National Institute of Information and Communications Technology (NICT), Suita, Osaka 565-0871, Japan

4. Author Contributions

I.Y., T.K., S.Nishida, H.H., K.M, K.I, and S.Nishimoto Designed Research;
I.Y., T.K., S.Nishida, H.H., and S.Nishimoto Performed Research;
I.Y., T.K., S.Nishida, H.H., K.M, K.I, and S.Nishimoto Analyzed data;
I.Y., T.K., and S.Nishida wrote the paper;

5. Correspondence should be addressed to:

Takatomi Kubo, takatomi-k@is.naist.jp, Graduate School of Science and Technology, Division of Information Science, Nara Institute of Science and Technology (NAIST), Ikoma, Nara 630-0192, Japan

6. **Number of Figures: 8**
7. **Number of Tables: 2**
8. **Number of Multimedia: 0**
9. **Number of words for Abstract: 144**
10. **Number of words for Significance Statement: 120**
11. **Number of words for Introduction: 746**
12. **Number of words for Discussion: 2507**

13. Acknowledgements

We thank Takao Nakagawa and Hidetake Uwano for helpful comments on the initial study design.

14. Conflict of Interest

The authors declare no competing financial interests.

15. Funding sources

This work was supported by JSPS KAKENHI Grant Number JP15H05311, JP16H05857, JP16H06569, JP17H01797, JP18K18108, JP18K18141, JP18J22957, and JST ERATO Grant Number JPMJER1801.

Expert programmers have fine-tuned cortical representations of source code

Abstract

Expertise enables humans to achieve outstanding performance on domain-specific tasks, and programming is no exception. Many studies 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 of programmers 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 a source-code categorization task. Our results suggest that programming expertise is built upon fine-tuned cortical representations specialized for the domain of programming.

Significance Statement

The expertise needed for programming has attracted increasing interest among researchers and educators in our computerized world. Many studies have demonstrated that expert programmers exhibit superior behavioral performance, knowledge structure, and selective attention; but how their brain accommodates such superiority is not well understood. In this paper we have recorded brain activities from subjects covering a wide range of programming expertise. The results show that functional categories of source code can be

decoded from their brain activity and the decoding accuracies on the seven brain regions in frontal, parietal, temporal cortices are significantly correlated with individual behavioral performances. This study provides evidence that outstanding performances of expert programmers are associated with domain-specific cortical representations in these widely distributed brain areas.

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 software projects and directly impact their success, this relatively new type of expertise is attracting increasing attention from modern industries (Li et al., 2015; Baltes and Diehl, 2018) and educational institutes (Heintz et al., 2016; Papavlasopoulou et al., 2018). Moreover, huge productivity variations were repeatedly found even between programmers with the same level of experience (Boehm and Papaccio, 1988; DeMarco and Lister, 2013). Previous studies have shown the psychological characteristics of expert programmers in their behaviors (Vessey, 1985; Koenemann and Robertson, 1991), knowledge structures (Fix et al., 1993; Von Mayrhauser and Vans, 1995), and eye movements (Uwano et al., 2006; Busjahn et al., 2015). 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). 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; including the middle frontal gyrus, inferior frontal gyrus, inferior parietal lobule, middle temporal gyrus (Siegmund et al., 2014, 2017). Their results suggested that program comprehension is associated with natural language processing, division of attention, and verbal/numerical working memory. Peitek *et al.* reanalyzed the same data as (Siegmund et al., 2014) to investigate the correlation between the BOLD activation strength and individual programming experience, which was determined by subject's self-estimation, but did not find any significant trend (Peitek et al., 2018a). An exploratory study argued that a correlation exists between activity pattern discriminability and subjects' grade point average (GPA) scores counting only

courses from the Computer Science department as a proxy for programming expertise (Floyd et al., 2017). However, the GPA scores would reflect a mixture of diverse factors (IQ, memory ability, calculation skills, etc.) and the assumed relationship to programming expertise was difficult to be empirically validated. Further, the main limitation of these prior studies is the use of a 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 the elucidation of the potential differences of brain functions related to the expertise.

Here we aim to identify the neural bases of programming expertise that contribute to the outstanding performances of expert programmers. To do this, we defined two fundamental factors in our experiment: An objectively determined reference of programming expertise and a laboratory task that exhibits experts' superior performances under the general constraints of fMRI experiments. First, we adopted the programmers' ratings in competitive programming contests (AtCoder, https://atcoder.jp/), which are objectively determined by the relative positions of their actual performances among thousands of programmers. We recruited top- and middle-rated programmers as well as novice controls to cover a wide range of programming expertise in our fMRI experiment. Second, we developed the program categorization task and confirmed that behavioral performances of this task were significantly correlated with the adopted reference of programming expertise. This confirmation allows us to expect an association between the outstanding performances of expert programmers and brain activity patterns recorded by fMRI while they performed this laboratory task.

Our core hypothesis is that higher programming expertise and experts' outstanding performances relate to specific multi-voxel pattern representations, potentially influenced by their domain-specific knowledge and training experiences. This hypothesis is 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 (de Borst et al., 2016; Martens et al., 2018; Gomez et al., 2019). For example, Bilalić *et al.* showed that the multi-voxel patterns in expert radiologists' Fusiform Face Area were more sensitive in differentiating X-ray images from control stimuli than novices (Bilalić et al., 2016). Similarly, identifying the multi-voxel pattern representations specific to expert programmers offers a good starting point for understanding the cognitive mechanisms behind programming expertise. From the previous studies on non-expert programmers and expertise in other domains, the high-level visual and left fronto-parietal regions might be inferred as potential neural correlates of programming expertise (Siegmund et al.,

2014; Bilalić, 2017). However, to the best of our knowledge, there is no prior evidence that directly associates programming expertise with specific brain regions. Thus, we employ a whole-brain searchlight analysis (Kriegeskorte et al., 2006) to identify the regions related to programming expertise.

Materials and Methods

Subjects

To begin this study, we defined three recruiting criteria: $Expert$, top 20% rankers in AtCoder who had an AtCoder rate equal to or higher than 1,200; $Middle$, 21-50% rankers who had an AtCoder rate between 500 and 1,199; $Novice$, subjects who had four years or less programming experience and no experience in competitive programming. We shared our recruiting message via mailing lists and messaging applications with diverse graduate or undergraduate student communities in Japan. Through this procedure, 95 programmers from 28 universities and three companies completed our entry questionnaire to be registered as candidate subjects. The list of candidate subjects consisted of 19 experts (all male), 43 middles (one female), and 33 novices (nine females). Nine left-handed subjects and 20 subjects with less than half a year experience in Java programming were excluded from the list. Five subjects aged under 20 years old were also excluded to avoid additional bureaucratic processes. We asked the remaining candidate subjects for experiment participation basically on first-in-first-out strategy. Note that setting $Novice$ as programmers who had an AtCoder rate under 500 was another potential recruiting criterion; but we did not adopt the criterion because low values in the rate reflects two indistinguishable factors: low programming expertise or not enough contest participation. In addition, possession of AtCoder rate itself could imply possession of moderate programming expertise. Thus, our recruiting criteria set $Novice$ as a programmer with shorter experience in programming and no experience in competitive programming.

Thirty healthy subjects (two females, aged between 20 and 24 years) with normal or corrected-to-normal vision participated in the experiment; see Table.1 for the demographic information of recruited subjects. All were right-handed (assessed by the Edinburgh Handedness Inventory (Oldfield, 1971), laterality quotient = 83.6 $\pm$ 24.0, ranged between +5.9 and +100) and understood basic Java grammars with at least half a year experience in Java programming. 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 based on the ranking on July 1 2017, respectively. Seven additional subjects were scanned but not included in the analysis because one (novice) showed neurological abnormality in MRI images, three (one expert and two middles) retired from the experiment without full completion, three (one expert and two novices) showed strongly-biased behavioral responses judged when the behavioral performance of one or more choices did not reach chance-level in the training experiments, signaling a strong response bias of sticking to a specific choice. This study was approved by the Ethics Committees of NAIST and CiNet and subjects gave written informed consent for participation. The sample size was chosen to match previous fMRI studies on human expertise with similar behavioral protocols (Amalric and Dehaene, 2016; Bilalić et al., 2016; de Borst et al., 2016).

Stimuli

For this study 72 code snippets written in Java were collected from an open codeset provided by AIZU ONLINE JUDGE (http://judge.u-aizu.ac.jp/onlinejudge/); an online judge system where many programming problems are listed and everyone can submit their own source code to answer those problems online. We selected four functional categories ($category$) and eleven subordinate concrete algorithms ($subcategory$) based on two popular textbooks about computer algorithms (Cormen et al., 2009; Sedgewick and Wayne, 2011); see Fig.1a and Extended Data Fig. 1-1 for the detailed descriptions. We first 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 and prior fMRI studies on programmers also used Java code snippets as experimental stimuli (Siegmund et al., 2014, 2017; Peitek et al., 2018a). To meet the screen size constraint in the MRI scanner, we excluded code snippets with a number of lines of more than 30 and a max number of characters per line of more than 120. From all remaining snippets, we created a set of 72 code snippets with minimum deviations of these numbers of lines and characters to minimize visual variation as experimental stimuli; the mean and standard deviation of the number of lines and max characters per line 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 the linear search class with twelve snippets (see Extended Data Fig. 1-2 for detailed statistics on each $category$ and $subcategory$ class). The

indentation styles of code snippets were normalized by replacing a tab-space with two white-spaces and user-defined functions were renamed to neutral such as "function1" because some of the functions indicated their algorithms explicitly (see Extended Data Fig. 1-3 for example snippets). 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 the undesirable effects of MRI signal instability. We used 72 code snippets as stimuli and each snippet was presented three times through the whole experiment (216 trials in total), but the same snippet appeared only once in a run. We employed PsychoPy (Peirce, 2007) (version 1.85.1) to display the code snippets in white text and a gray background without syntax highlighting to minimize visual variations. In each trial of the program categorization tasks, a Java code snippet was displayed for ten seconds after a fixation-cross presentation for two seconds. Subjects then responded within four seconds by pressing buttons placed under the right hand to indicate which $category$ class was most plausible for the code snippet and all response data were automatically collected for the calculation of individual behavioral performance. To clarify classification criteria, a brief explanation about each $category$ class was provided before the experiment started. The presentation order of the code snippets was pseudo-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.

All subjects took two additional sessions, named "Training" and "Post-MRI", outside of the MRI scanner using a laptop computer and PsychoPy to display source code stimuli. The training session was performed within ten days before the fMRI experiment to mitigate potential confounds caused by task unfamiliarity. The session consisted of three separate runs with the same program categorization task as the fMRI experiment. A different set of 72 Java code snippets from those used in the MRI experiment, which covered the same algorithms, was used as stimuli in the training session; each snippet was presented once or twice in the entire session but the same snippet did not appear twice in a run. The post-MRI session was performed

within ten days after the fMRI experiment for assessment of individual ability in $subcategory$ categorizations and was consisted of two separate runs using the same codeset as the fMRI experiment. Before the post-MRI session started, we explained the existence of $subcategory$ to the subjects and assessed whether they recognized subcategory classes during the fMRI experiment using a questionnaire. Program categorization tasks in the post-MRI session followed the same procedure as the fMRI session. In each trial, a Java code snippet was displayed for ten seconds after a fixation-cross presentation for two seconds. Then, within four seconds, the subjects were asked to classify the given code snippet from two or three choices of $subcategory$ classes according to its superordinate category, e.g., "bubble sort", "insertion sort", and "selection sort" were displayed when the snippet in "sort" category was presented. We calculated behavioral performance as the ratio of correct-answer-trials to all-trials; unanswered trials, i.e. no button input within the response phase, were regarded as "incorrect" for this calculation. Chance-level behavioral performance was 25% in the training sessions and fMRI experiments and 37.25% in the post-MRI sessions adjusted for imbalanced numbers of answer choices. Again, these two additional sessions were performed outside of the MRI scanner, in other words, every subject did only one experiment with fMRI scanning.

MRI data acquisition

MRI data were collected using a 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 $\times$ 192 mm^2, slice thickness = 2 mm, slice gap = 0 mm, voxel size = $2 \times 2 \times 2.01$ mm^3, multi-band factor = 3). A T1-weighted magnetization-prepared rapid acquisition with a 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 $\times$ 256 mm^2, slice thickness = 1 mm, slice gap = 0 mm, voxel size = $1 \times 1 \times 1$ mm^3).

MRI data preprocessing

We used the Statistical Parametric 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 Institute 152-brain average template space and resampled to a voxel size of $2 \times 2 \times 2$ mm^3. MRI signals at each voxel were high-pass2013filtered 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 (Ridgway et al., 2009). 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.

Multi-voxel pattern analysis

We used whole-brain searchlight analysis (Kriegeskorte et al., 2006) to examine where significant decoding accuracies exist using the Decoding Toolbox (Hebart et al., 2015) (version 3.99) and LIBSVM (Chang and Lin, 2011) (version 3.17). 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. A linear-kernel support vector machine (SVM) classifier was trained and evaluated using a leave-one-run-out cross-validation procedure, which iteratively treated data in a single run for testing and the 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. The SVM classifier was trained with three cost parameter candidates [0.1, 1, 10], which control the tradeoff between margin maximization and the tolerance of misclassification rate in the training step, and the best parameter was chosen by a grid search in nested cross-validations. The outlier boundary and cost parameter candidates were selected based on the estimated computational load and the documents of tools employed. Specifically, We here adopted a relatively small

set of parameter candidates due to the constraint of the high computational load of searchlight analysis. 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 procedures were performed independently for each subject and a whole-brain decoding accuracy map was obtained per subject. We then conducted second-level analyses to examine the significance of decoding accuracies and the correlations between individual decoding accuracies and behavioral performances. For this purpose, the 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 as implemented in SPM12. The analysis tested the significance of group-level decoding accuracy and Pearson's correlation coefficient between individual decoding accuracies and behavioral performances. A relatively strict statistical threshold of voxel-level $p < 0.05$ FWE-corrected was used for decoding accuracy tests and a standard threshold of voxel-level $p < 0.001$ uncorrected and cluster-level $p < 0.05$ FWE-corrected was used for correlation tests. The chance-level accuracy (25% in *category* decoding and 9.72% in *subcategory* decoding; adjusted for imbalanced numbers of exemplar) and zero correlation were adopted as null hypotheses. Additionally, the resultant significant searchlight maps, i.e. decoding accuracy map and correlation map to behavioral performance, were superimposed on a single cortical surface of the ICBM152 template brain using BrainNet viewer (Xia et al., 2013). We performed this superimposition to identify the searchlight centers that had both sufficient information to represent functional categories of source code and significant correlation between individual behavioral performances and decoding accuracies.

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.

Results

Behavioral data

We evaluated the relationship between the adopted reference of programming expertise and behavioral performance on the program categorization task. A significant correlation was observed between AtCoder rate (M = 954.3, SD = 864.6) and behavioral performance in the fMRI experiments (M = 76.0, SD = 13.5 [%]), r = 0.593, p = 0.0059, n = 20 (Fig.2a). The correlation remained significant if we included behavioral performances of non-rate-holders (i.e. novices) as zero-rated subjects; r = 0.722, p = 0.000007, n = 30. We additionally found a positive correlation between AtCoder rate and behavioral performance on subcategory categorization in the post-MRI experiments (M = 65.9, SD = 17.0 [%]), r = 0.688, p = 0.0008, n = 20 (Fig.2b). The significant correlation also remained significant if we included non-rate-holder subjects; r = 0.735, p = 0.000004, n = 30. From all behavioral data, we certainly concluded that behavioral performances on the program categorization task significantly correlated with expertise of competitive programming. The behavioral evidence allowed us to study the potential association between experts' outstanding performances and brain activity patterns measured using fMRI while subjects performed this laboratory task.

Mutli-voxel activity patterns associated with programming expertise

We first examined where we could decode the functional categories of source code from programmers' brain activity. Fig.3 visualizes the searchlight centers that showed significantly higher decoding accuracy than chance as estimated from all subject data using a relatively strict whole-brain statistical threshold (voxel-level $p < 0.05$ FWE-corrected). The figure shows that significant decoding accuracies were observed in the broad areas of the 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 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.

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.4a 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.4b and Table 2 for the list of significant clusters. In this correlation analysis, the right IFG Tri showed the highest peak correlation coefficient. 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 behavioral 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.3.

Two previous analyses separately showed where significant decoding accuracies exist and whether the 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. Specifically, the two significant searchlight maps, i.e. decoding accuracy map and correlation map to behavioral performance, were superimposed on a single cortical surface to investigate the overlap between them. 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.5a. 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) as shown in Fig.5b. These results revealed a tight association between superior behavioral performances of expert programmers and 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 activity to examine finer-level cortical representations. In our experiment, subjects responded 'sort' when

they had 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 behavioral performances in subordinate-level categorizations as well as basic-level categorizations (Tanaka and Taylor, 1991). In fact, we have observed that the ability to differentiate $subcategory$ classes significantly correlated to programming expertise in competitive programming (Fig.2b). This observation implies that the detailed difference of source code functionalities might be represented in programmers' brain activity patterns. 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 the 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.6 illustrates the searchlight centers that showed significantly higher $subcategory$ decoding accuracy than chance (9.72%; corrected for imbalanced exemplars) using a threshold of voxel-level $p < 0.05$ FWE-corrected. The 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.7). 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. 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.8a. 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. Decoding accuracies on the bilateral inferior frontal gyrus pars triangularis, left inferior parietal lobule, left supramarginal gyrus, left middle and inferior temporal gyri, and right middle frontal gyrus were significantly correlated with individual behavioral performances on the program categorization task. Furthermore, decoding accuracies of subcategory on the left supramarginal and superior temporal gyri were also strongly correlated with the behavioral performances while the subordinate-level representations were not directly induced by the performing tasks. Our results revealed an association between the outstanding performances of expert programmers 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 (Siegmund et al., 2014, 2017; Floyd et al., 2017; Castelhano et al., 2019; Peitek et al., 2018a,b). Exceptionally, an exploratory study reported that BOLD signal discriminability between code and text comprehension was negatively correlated with participants' GPA scores in a university (Floyd et al., 2017). 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 to expert programmers' outstanding performances. To address the goal, we adopted an objectively-determined reference of programming expertise and recruited a population of subjects covering a wide range of programming expertise. 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 (Siegmund et al., 2014, 2017; Peitek et al., 2018a). 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 may suggest that both program comprehension processes and its related expertise depend on the same set of brain regions.

The potential roles of the specified brain regions in our study should be addressed to orient future researches on programming activity and expertise. First, the left IFG Tri and the left posterior MTG are

frequently involved in semantic selecting/retrieving tasks (Demonet et al., 1992; Thompson-Schill et al., 1997; Simmons et al., 2005; Price, 2012). Several studies indicated that these two regions are sensitive to cognitive demands for directing semantic knowledge retrieval in a goal-oriented way (Rodd et al., 2005; Kuhl et al., 2007; Whitney et al., 2010). The involvement of the two regions in our findings may be induced by similar demands specialized for the retrieval of program functional categories and suggest that higher programming expertise is related to the abilities of goal-oriented knowledge retrieval. Second, many neuroscientists have shown the left IPL and SMG to be functionally related to visual word reading (Bookheimer et al., 1995; Philipose et al., 2007; Stoeckel et al., 2009) and episodic memory retrieval (Wagner et al., 2005; Vilberg and Rugg, 2008; O'Connor et al., 2010). 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. In addition, the set of IFG and IPL has been frequently discussed together as a fronto-parietal network and they often show synchronous activity in a wide range of tasks (Watson and Chatterjee, 2012; Ptak et al., 2017). Importantly, a recent fMRI study on programmers suggested an association between program comprehension and fronto-parietal network that was functionally related to formal logical inference (Liu et al., 2020). Our results are consistent with these findings, implying that the fronto-parietal network plays a key role in experts' program comprehension processes.

Other novel findings in the present study included 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 of non-expert subjects (Siegmund et al., 2014, 2017; Floyd et al., 2017; Peitek et al., 2018a), suggesting that the regions might be more related to expert programmers' 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 (Chelazzi et al., 1993; Nobre et al., 1994; Kriegeskorte et al., 2008), our results may suggest that the high-level visual cortex in expert programmers could be fine-tuned by their training experience to re-

alize faster program comprehension process. From another perspective, the observed map involving the left IFG Tri, IPL, and MTG/IT (Fig.4a) could be associated with a semantic system in the brain (Patterson et al., 2007; Binder et al., 2009). Our results might suggest that an expert programmer's brain recruits a similar language-related network for both natural language processing and program comprehension. In contrast, the primary visual area showed significant decoding accuracy but no correlation to programming expertise. The primary visual area mainly reflects primitive visual features such as color, contrast, spatial frequency (Tong, 2003) while computations in the high-level visual cortex are characterized by both bottom-up (i.e. how stimuli are visually represented) and top-down (how the representation is used for a cognitive task) effects (Kay and Yeatman, 2017). Previous studies indicated that fine-tuned representations in the high-level visual cortex, rather than in the primary visual area, could be associated with visual expertise (Bilalić et al., 2016) and reading skill (Kubota et al., 2019). In our experiment, the primary visual area represented a large amount of visual information regardless of programming expertise levels because all subjects were presented with the same set of code snippets inducing similar visual patterns on their retinas. Therefore, the information in the primary visual area was sufficient to decode $category$ and $subcategory$ classes but the decoding accuracies were not necessarily to be correlated with individual behavioral performances. Meanwhile, the amount of information represented in the high-level visual cortex might be modulated by individual programming expertise. In line with previous expertise studies, our results imply that expertise in program comprehension could be mainly associated with high-level visual perception.

The right MFG and IFG Tri are functionally related to stimulus-driven attention control (Corbetta et al., 2008; Japee et al., 2015). 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 (Bilalić et al., 2011) and abacus experts (Tanaka et al., 2002; Hanakawa et al., 2003) showed additional right hemisphere region involvements when performing their domain-specific tasks. Several fMRI studies further suggest that such activation shifts from left to right hemisphere may be related to experts' cognitive strategy changes (Bilalić et al., 2011; Tanaka et al., 2012). 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 comprehen-

sion, which becomes feasible by experts' domain-specific knowledge (Koenemann and Robertson, 1991; Fix et al., 1993; Von Mayrhauser and Vans, 1995). 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. From another perspective, activations in the prefrontal and parietal regions including bilateral IFG/MFG and left IPL have been associated with the extent of cognitive demands (Harvey et al., 2005). While our study did not have a direct indicator of cognitive demands across categories, the difference in behavioral performances for each $category$ can be a clue to assess the extent of cognitive demand across the categories. We used the one-way ANOVA to test the difference in mean behavioral performances between categories but no significant difference was found for any groupings (see Extended Data Fig. 2-1). Although these results do not provide a direct indication of cognitive demands across categories, we have no positive evidence that the extent of cognitive demands had a significant effect on the observed decoding accuracies.

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. These two regions are functionally related to pre-lexical and phonological processing in natural language comprehension (Demonet et al., 1992; Moore and Price, 1999; Burton et al., 2001). 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 (Rodd et al., 2005; Whitney et al., 2010; Price, 2012). 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. Further, Mkrtychian *et al.* have associated STG, MTG, and IFG with the processing of abstract concepts in their review on concreteness effects (Mkrtychian et al., 2019), implying that representations in these three regions could reflect relative differences in abstractness between the $category$ and $subcategory$ in our study. These interpretations might suggest a hypothesis that an expert programmer's brain has a hierarchical semantic processing system to obtain mental representations of source code for multiple levels of abstraction.

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 (Diedrichsen and Kriegeskorte, 2017). Several researchers have demonstrated frameworks to decode arbitrary objects using a set of computational visual features representing categories of target objects (Horikawa and Kamitani, 2017) and to decode perceptual experiences evoked by natural movies using word-based distributed representations (Nishida and Nishimoto, 2018). Other studies have also used word-based distributed representations to systematically map semantic selectivity across the cortex (Huth et al., 2016; Pereira et al., 2018). Meanwhile, researchers in the program analysis domain have proposed distributed representations of source code based on abstract syntax tree (AST) (Alon et al., 2019a; Zhang et al., 2019). Alon *et al.*, for instance, have presented continuous distributed vectors representing the functionality of source code using AST and path-attention neural network (Alon et al., 2019b). 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 connects semantic features of source code to programmers' perceptual experiences.

Limitations of the study

The results obtained via the present study were limited to a specific type of programming expertise evaluated by the expertise reference 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. Perhaps there is a qualitative gap between expertise in competitive programming and practical/industrial software development. For example, 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 such type of programming expertise. The program categorization task used in this study primarily evaluated the skill in recognizing algorithms quickly and accurately, which was one aspect of a wide range of cognitive skills that constitute programming expertise. We considered that the evaluated skill is related to program comprehension and is also connected to skills in code refactoring and debugging because these processes require a deep understanding of algorithms or how the code works; while its relation to writing code is not assessed in this study. 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 objectively-determined reference 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 for future researches to investigate the neural bases of programming expertise and related abilities.

Our experiment, which was designed to fit the general constraints of fMRI measurement, might embrace several caveats to external validity. First, we used the relatively small code snippets with 30 lines at maximum due to the constraint of the MRI screen size. Behavioral performances on system-level source code were not assessed in the study. Thus, generalizing our results to the expertise in systems-level program comprehension was not guaranteed. Second, only Java code snippets were used as experimental stimuli in this study. The results obtained via the experiments might be biased by the programming language selected; for example, Python has more natural-language-like syntax than Java and might induce more activation in language-related brain regions. While a recent fMRI study has examined brain activities elicited by code written in two programming languages (Python and ScratchJr) (Ivanova et al., 2020), it is still unclear whether the choice of a specific programming language can alter an expert's brain activity pattern. The relationship between programming expertise and types of programming languages (e.g. procedural vs. functional languages) is expected to be examined in future work.

Another potential concern of the present study was the unfair gender balance in the subject population. While 95 programmers completed our entry questionnaire to be registered as candidate subjects, only one middle-level woman candidate and zero woman expert were found (see Subjects section in Materials and Methods). From this situation, we recognized the unavoidable gender bias in our target population. To properly cover a wide range of programming expertise, we were forced to give up on maintaining gender balance at each expertise level. However, several fMRI studies have reported possible gender differences in behavior, cognitive function, and neuroimaging data (David et al., 2018; Huang et al., 2020). The results obtained via this study might be biased by the gender imbalance of the subject population. Future work

should investigate whether behavioral and cognitive differences would be found between man and woman programmers. In addition, while our sample size was determined in line with previous expertise studies, ten subjects for each expertise level was not a big population and were insufficient to show statistically significant results between different expertise classes. Therefore, making mention of comparison between novice-middle or middle-expert must be with great caution. Larger samples would be desirable in future replication or follow-up studies.

Conclusion

Our findings reveal an association between programming expertise and cortical representations of program source code in a programmer's brain. We demonstrated that functional categories of source code can be decoded from programmer's brain activity and the decoding accuracies on the seven regions in the frontal, parietal, and temporal cortices were significantly correlated with individual behavioral performances. The results additionally suggest that cortical representations of fine functional categories ($subcategory$) on the left SMG and STG might be associated with advanced-level programming expertise. Although research on the neural basis of programming expertise is still in its infancy, we believe that our study extends the existing human expertise literature into the domain of programming by demonstrating that top-level programmers have domain-specific cortical representations.

References

Uri Alon, Omer Levy, and Eran Yahav. code2seq: Generating sequences from structured representations of code. *Proceedings of the 7th International Conference on Learning Representations*, 2019a.

Uri Alon, Meital Zilberstein, Omer Levy, and Eran Yahav. Code2vec: Learning distributed representations of code. *Proceedings of the 46th ACM SIGPLAN Symposium on Principles of Programming Languages*, 2019b.

Marie Amalric and Stanislas Dehaene. Origins of the brain networks for advanced mathematics in expert mathematicians. *Proceedings of the National Academy of Sciences*, 113(18):4909–4917, 2016.

Sebastian Baltes and Stephan Diehl. Towards a theory of software development expertise. *Proceedings of the ACM 26th Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 187–200, 2018.

Merim Bilalić. *The Neuroscience of Expertise*. Cambridge Fundamentals of Neuroscience in Psychology. Cambridge University Press, 2017.

Merim Bilalić, Andrea Kiesel, Carsten Pohl, Michael Erb, and Wolfgang Grodd. It takes two–skilled recognition of objects engages lateral areas in both hemispheres. *PLoS One*, 6(1):e16202, 2011.

Merim Bilalić, Thomas Grottenthaler, Thomas Nägele, and Tobias Lindig. The faces in radiological images: fusiform face area supports radiological expertise. *Cerebral Cortex*, 26(3):1004–1014, 2016.

Jeffrey R Binder, Rutvik H Desai, William W Graves, and Lisa L Conant. Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, 19(12): 2767–2796, 2009.

Barry W Boehm and Philip N. Papaccio. Understanding and controlling software costs. *IEEE Transactions on Software Engineering*, 14(10):1462–1477, 1988.

Susan Y Bookheimer, Thomas A Zeffiro, Teresa Blaxton, William Gaillard, and William Theodore. Regional cerebral blood flow during object naming and word reading. *Human Brain Mapping*, 3(2):93–106, 1995.

Martha W Burton, Douglas C Noll, and Steven L Small. The anatomy of auditory word processing: individual variability. *Brain and language*, 77(1):119–131, 2001.

Teresa Busjahn, Roman Bednarik, Andrew Begel, Martha Crosby, James H Paterson, Carsten Schulte, Bonita Sharif, and Sascha Tamm. Eye movements in code reading: Relaxing the linear order. *Proceedings of the IEEE 23rd International Conference on Program Comprehension*, pages 255–265, 2015.

Joao Castelhano, Isabel C Duarte, Carlos Ferreira, Joao Duraes, Henrique Madeira, and Miguel Castelo-Branco. The role of the insula in intuitive expert bug detection in computer code: an fMRI study. *Brain Imaging and Behavior*, 13(3):623–637, 2019.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3), 2011.

Leonardo Chelazzi, Earl K Miller, John Duncan, and Robert Desimone. A neural basis for visual search in inferior temporal cortex. *Nature*, 363(6427):345–347, 1993.

Maurizio Corbetta, Gaurav Patel, and Gordon L Shulman. The reorienting system of the human brain: from environment to theory of mind. *Neuron*, 58(3):306–324, 2008.

Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms, Third Edition*. The MIT Press, 2009. ISBN 0262033844, 9780262033848.

Sean P David, Florian Naudet, Jennifer Laude, Joaquim Radua, Paolo Fusar-Poli, Isabella Chu, Marcia L Stefanick, and John PA Ioannidis. Potential reporting bias in neuroimaging studies of sex differences. *Scientific reports*, 8(1):1–8, 2018.

Aline W de Borst, Giancarlo Valente, Iiro P Jääskeläinen, and Pia Tikka. Brain-based decoding of mentally imagined film clips and sounds reveals experience-based information patterns in film professionals. *NeuroImage*, 129:428–438, 2016.

Tom DeMarco and Tim Lister. *Peopleware: Productive Projects and Teams, Third Edition*. Addison-Wesley Professional, 2013. ISBN 0321934113, 9780321934116.

Jean-François Demonet, Francois Chollet, Stuart Ramsay, Dominique Cardebat, Jean-Luc Nespoulous, Richard Wise, André Rascol, and Richard Frackowiak. The anatomy of phonological and semantic processing in normal subjects. *Brain*, 115(6):1753–1768, 1992.

Jörn Diedrichsen and Nikolaus Kriegeskorte. Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS computational biology*, 13(4):e1005508, 2017.

Vikki Fix, Susan Wiedenbeck, and Jean Scholtz. Mental representations of programs by novices and experts. *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, pages 74–79, 1993.

Benjamin Floyd, Tyler Santander, and Westley Weimer. Decoding the representation of code in the brain: An fMRI study of code review and expertise. *Proceedings of the IEEE/ACM 39th International Conference on Software Engineering*, pages 175–186, 2017.

Jesse Gomez, Michael Barnett, and Kalanit Grill-Spector. Extensive childhood experience with pokémon suggests eccentricity drives organization of visual cortex. *Nature Human Behaviour*, 3(6):611, 2019.

Takashi Hanakawa, Manabu Honda, Tomohisa Okada, Hidenao Fukuyama, and Hiroshi Shibasaki. Neural correlates underlying mental calculation in abacus experts: a functional magnetic resonance imaging study. *NeuroImage*, 19(2):296–307, 2003.

Philippe-Olivier Harvey, Philippe Fossati, Jean-Baptiste Pochon, Richard Levy, Guillaume LeBastard, Stéphane Lehéricy, Jean-François Allilaire, and Bruno Dubois. Cognitive control and brain resources in major depression: an fmri study using the n-back task. *Neuroimage*, 26(3):860–869, 2005.

Martin N Hebart, Kai Görgen, and John-Dylan Haynes. The decoding toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. *Frontiers in Neuroinformatics*, 8:88, 2015.

Fredrik Heintz, Linda Mannila, and Tommy Färnqvist. A review of models for introducing computational thinking, computer science and computing in K-12 education. *Proceedings of the IEEE 46th Frontiers in Education conference*, pages 1–9, 2016.

Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, 8:15037, 2017.

Yu Huang, Kevin Leach, Zohreh Sharafi, Nicholas McKay, Tyler Santander, and Westley Weimer. Biases and differences in code review using medical imaging and eye-tracking: Genders, humans, and machines. *Proceedings of the ACM 28th Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2020.

Alexander G Huth, Wendy A de Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453, 2016.

Anna A Ivanova, Shashank Srikant, Yotaro Sueoka, Hope H Kean, Riva Dhamala, Una-May O'reilly, Ma-

rina U Bers, and Evelina Fedorenko. Comprehension of computer code relies primarily on domain-general executive resources. *BioRxiv*, 2020.

Shruti Japee, Kelsey Holiday, Maureen D Satyshur, Ikuko Mukai, and Leslie G Ungerleider. A role of right middle frontal gyrus in reorienting of attention: a case study. *Frontiers in Systems Neuroscience*, 9:23, 2015.

Kendrick N Kay and Jason D Yeatman. Bottom-up and top-down computations in word-and face-selective cortex. *Elife*, 6:e22341, 2017.

Jürgen Koenemann and Scott P Robertson. Expert problem solving strategies for program comprehension. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 125–130, 1991.

Nikolaus Kriegeskorte, Rainer Goebel, and Peter Bandettini. Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, 103(10):3863–3868, 2006.

Nikolaus Kriegeskorte, Marieke Mur, Douglas A Ruff, Roozbeh Kiani, Jerzy Bodurka, Hossein Esteky, Keiji Tanaka, and Peter A Bandettini. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60(6):1126–1141, 2008.

Emily C Kubota, Sung Jun Joo, Elizabeth Huber, and Jason D Yeatman. Word selectivity in high-level visual cortex and reading skill. *Developmental Cognitive Neuroscience*, 36:100593, 2019.

Brice A Kuhl, Nicole M Dudukovic, Itamar Kahn, and Anthony D Wagner. Decreased demands on cognitive control reveal the neural processing benefits of forgetting. *Nature Neuroscience*, 10(7):908, 2007.

Paul Luo Li, Andrew J Ko, and Jiamin Zhu. What makes a great software engineer? *Proceedings of the IEEE/ACM 37th IEEE International Conference on Software Engineering*, pages 700–710, 2015.

Yun-Fei Liu, Judy Kim, Colin Wilson, and Marina Bedny. Computer code comprehension shares neural resources with formal logical inference in the fronto-parietal network. *BioRxiv*, 2020.

Farah Martens, Jessica Bulthé, Christine van Vliet, and Hans Op de Beeck. Domain-general and domain-specific neural changes underlying visual expertise. *NeuroImage*, 169:80–93, 2018.

Nadezhda Mkrtychian, Evgeny Blagovechtchenski, Diana Kurmakaeva, Daria Gnedykh, Svetlana Kostro-

mina, and Yury Shtyrov. Concrete vs. abstract semantics: from mental representations to functional brain mapping. *Frontiers in human neuroscience*, 13:267, 2019.

Caroline J Moore and Cathy J Price. Three distinct ventral occipitotemporal regions for reading and object naming. *NeuroImage*, 10(2):181–192, 1999.

Satoshi Nishida and Shinji Nishimoto. Decoding naturalistic experiences from human brain activity via distributed representations of words. *NeuroImage*, 180:232–242, 2018.

Anna C Nobre, Truett Allison, Gregory McCarthy, et al. Word recognition in the human inferior temporal lobe. *Nature*, 372(6503):260–263, 1994.

Akira R O'Connor, Sanghoon Han, and Ian G Dobbins. The inferior parietal lobule and recognition memory: expectancy violation or successful retrieval? *Journal of Neuroscience*, 30(8):2924–2934, 2010.

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

Sofia Papavlasopoulou, Kshitij Sharma, and Michail N Giannakos. How do you feel about learning to code? investigating the effect of children's attitudes towards coding using eye-tracking. *International Journal of Child-Computer Interaction*, 17:50–60, 2018.

Karalyn Patterson, Peter J Nestor, and Timothy T Rogers. Where do you know what you know? the representation of semantic knowledge in the human brain. *Nature reviews neuroscience*, 8(12):976–987, 2007.

Jonathan W Peirce. Psychopy—psychophysics software in python. *Journal of Neuroscience Methods*, 162: 8–13, 2007.

Norman Peitek, Janet Siegmund, Sven Apel, Christian Kästner, Chris Parnin, Anja Bethmann, Thomas Leich, Gunter Saake, and André Brechmann. A look into programmers' heads. *IEEE Transactions on Software Engineering*, 46(4):442–462, 2018a.

Norman Peitek, Janet Siegmund, Chris Parnin, Sven Apel, Johannes C. Hofmeister, and André Brechmann. Simultaneous measurement of program comprehension with fMRI and eye tracking: A case study.

Proceedings of the ACM/IEEE 12th International Symposium on Empirical Software Engineering and Measurement, 2018b.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, 9(1):963, 2018.

Lisa E Philipose, Rebecca F Gottesman, Melissa Newhart, Jonathan T Kleinman, Edward H Herskovits, Mikolaj A Pawlak, Elisabeth B Marsh, Cameron Davis, Jennifer Heidler-Gary, and Argye E Hillis. Neural regions essential for reading and spelling of words and pseudowords. *Annals of Neurology : Official Journal of the American Neurological Association and the Child Neurology Society*, 62(5):481–492, 2007.

Cathy J Price. A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *NeuroImage*, 62(2):816–847, 2012.

Radek Ptak, Armin Schnider, and Julia Fellrath. The dorsal frontoparietal network: a core system for emulated action. *Trends in cognitive sciences*, 21(8):589–599, 2017.

Gerard R Ridgway, Rohani Omar, Sébastien Ourselin, Derek LG Hill, Jason D Warren, and Nick C Fox. Issues with threshold masking in voxel-based morphometry of atrophied brains. *NeuroImage*, 44(1): 99–111, 2009.

Jennifer M Rodd, Matthew H Davis, and Ingrid S Johnsrude. The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity. *Cerebral Cortex*, 15(8):1261–1269, 2005.

Edmund T Rolls, Marc Joliot, and Nathalie Tzourio-Mazoyer. Implementation of a new parcellation of the orbitofrontal cortex in the automated anatomical labeling atlas. *Neuroimage*, 122:1–5, 2015.

Robert Sedgewick and Kevin Wayne. *Algorithms, Forth Edition*. Addison-Wesley Professional, 2011. ISBN 032157351X, 9780321573513.

Janet Siegmund, Christian Kästner, Sven Apel, Chris Parnin, Anja Bethmann, Thomas Leich, Gunter Saake, and André Brechmann. Understanding understanding source code with functional magnetic resonance

imaging. *Proceedings of the IEEE/ACM 36th International Conference on Software Engineering*, pages 378–389, 2014.

Janet Siegmund, Norman Peitek, Chris Parnin, Sven Apel, Johannes Hofmeister, Christian Kästner, Andrew Begel, Anja Bethmann, and André Brechmann. Measuring neural efficiency of program comprehension. *Proceedings of the 11th Joint Meeting on Foundations of Software Engineering*, pages 140–150, 2017.

Alan Simmons, Daniel Miller, Justin S Feinstein, Terry E Goldberg, and Martin P Paulus. Left inferior prefrontal cortex activation during a semantic decision-making task predicts the degree of semantic organization. *NeuroImage*, 28:30–38, 2005.

Cornelia Stoeckel, Patricia M Gough, Kate E Watkins, and Joseph T Devlin. Supramarginal gyrus involvement in visual word recognition. *Cortex*, 45(9):1091–1096, 2009.

James W Tanaka and Marjorie Taylor. Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, 23(3):457–482, 1991.

Satoshi Tanaka, Chikashi Michimata, Tatsuro Kaminaga, Manabu Honda, and Norihiro Sadato. Superior digit memory of abacus experts: an event-related functional MRI study. *NeuroReport*, 13(17):2187–2191, 2002.

Satoshi Tanaka, Keiko Seki, Takashi Hanakawa, Madoka Harada, Sho K Sugawara, Norihiro Sadato, Katsumi Watanabe, and Manabu Honda. Abacus in the brain: a longitudinal functional MRI study of a skilled abacus user with a right hemispheric lesion. *Frontiers in Psychology*, 3:315, 2012.

Sharon L Thompson-Schill, Mark D'Esposito, Geoffrey K Aguirre, and Martha J Farah. Role of left inferior prefrontal cortex in retrieval of semantic knowledge: a reevaluation. *Proceedings of the National Academy of Sciences*, 94(26):14792–14797, 1997.

Frank Tong. Primary visual cortex and visual awareness. *Nature Reviews Neuroscience*, 4(3):219–229, 2003.

Hidetake Uwano, Masahide Nakamura, Akito Monden, and Kenichi Matsumoto. Analyzing individual

performance of source code review using reviewers' eye movement. *Proceedings of the Eye tracking Research and Applications Symposium*, pages 133–140, 2006.

Iris Vessey. Expertise in debugging computer programs: A process analysis. *International Journal of Man-Machine Studies*, 23(5):459–494, 1985.

Kaia L Vilberg and Michael D Rugg. Memory retrieval and the parietal cortex: a review of evidence from a dual-process perspective. *Neuropsychologia*, 46(7):1787–1799, 2008.

Anneliese Von Mayrhauser and A Marie Vans. Program comprehension during software maintenance and evolution. *Computer*, 28(8):44–55, 1995.

Anthony D Wagner, Benjamin J Shannon, Itamar Kahn, and Randy L Buckner. Parietal lobe contributions to episodic memory retrieval. *Trends in Cognitive Sciences*, 9(9):445–453, 2005.

Christine E Watson and Anjan Chatterjee. A bilateral frontoparietal network underlies visuospatial analogical reasoning. *Neuroimage*, 59(3):2831–2838, 2012.

Carin Whitney, Elizabeth Jefferies, and Tilo Kircher. Heterogeneity of the left temporal lobe in semantic representation and control: priming multiple versus single meanings of ambiguous words. *Cerebral Cortex*, 21(4):831–844, 2010.

Mingrui Xia, Jinhui Wang, and Yong He. BrainNet Viewer: a network visualization tool for human brain connectomics. *PLoS One*, 8(7):e68910, 2013.

Jean Zhang, Xu Wang, Hongyu Zhang, Hailong Sun, Kaixuan Wang, and Xudong Liu. A novel neural source code representation based on abstract syntax tree. *Proceedings of the IEEE/ACM 41th International Conference on Software Engineering*, pages 783–794, 2019.

Legends

Extended Data Figure 1-1: Description of categories and subcategories provided to subjects in the experiment.

Extended Data Figure 1-2: Statistics of Java code snippets for each category and subcategory class. Numerics from 3rd (LOC) to last columns denote 'MEAN±SD'. Abbreviations: LOC: Lines of code, CPL: Max number of characters per line.

Extended Data Figure 1-3: Java code snippets used in the study. 72 types of Java code snippet were used in this study. Each belonged to one subcategory and its corresponding category shown in Figure.1a. This figure shows example snippets for each subcategory class.

Extended Data Figure 2-1: Behavioral performance of each category in the fMRI experiment. Numerics from 3rd (Math) to last columns denote 'MEAN±SD'. One-way ANOVA found no significant difference in behavioral performances between categories for any groupings (Expert, $F(3,36) = 1.38$, $p = 0.27$; Middle, $F(3,36) = 2.99$, $p = 0.06$; Novice, $F(3,36) = 2.81$, $p = 0.07$; All, $F(3,116) = 2.02$, $p = 0.12$).

Extended Data Figure 3-1: Box plots of the voxel-level peak category decoding accuracies on several brain regions. Each dot represents decoding accuracy of individual subject. The dashed line indicates chance-level accuracy (25%). Abbreviations: SMG, Supramarginal gyrus; IPL, Inferior parietal lobule; MTG, Middle temporal gyrus; IT, Inferior temporal gyrus; IFG Tri, Inferior frontal gyrus pars triangularis.

Extended Data Figure 6-1: Box plots of the voxel-level peak subcategory decoding accuracies on several brain regions. Each dot represents decoding accuracy of individual subject. The dashed line indicates chance-level accuracy (9.72%). Abbreviations: SMG, Supramarginal gyrus; IPL, Inferior parietal lobule; MTG, Middle temporal gyrus; IT, Inferior temporal gyrus; IFG Tri, Inferior frontal gyrus pars triangularis.

Custom software code: The custom MATLAB code used for the decoding analysis in the paper.

Figure, Table, Visual Abstract, and Multimedia

Table 1: **Demographic information of recruited subjects.** Numerics from 4th (Age) to last columns denote 'MEAN ± SD'. Abbreviations: PE, programming experience; JE, Java experience; CPE, competitive programming experience. Significant differences were observed between PE of Expert - Novice, Middle - Novice; CPE of Expert - Middle (two-sample t-test, $p < 0.05$ FDR-corrected).

	N	Sex (M/F)	Age	AtCoder rate	PE (year)	JE (year)	CPE (year)
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

Table 2: **Clusters showing significant correlations between behavioral performance and category decoding accuracy** (voxel-level $p < 0.001$ and cluster-level $p < 0.05$, FWE-corrected). Region names were identified using Automated anatomical labelling atlas 2 (Rolls et al., 2015).

	MNI coordinates					
Region name	X	Y	Z	Peak corr. (r)	T-value	Cluster extent
R IFG (p. Triangularis)	46	22	8	0.789	6.81	369
L Posterior-Medial Frontal	-12	0	66	0.711	5.36	298
R Superior Medial Gyrus	6	52	42	0.699	5.17	587
L Inferior Parietal Lobule	-56	-28	50	0.698	5.16	649
R Superior Frontal Gyrus	24	4	60	0.675	4.84	428
L IFG (p. Triangularis)	-52	30	24	0.671	4.79	346
L Inferior Temporal Gyrus	-50	-54	0	0.635	4.35	347

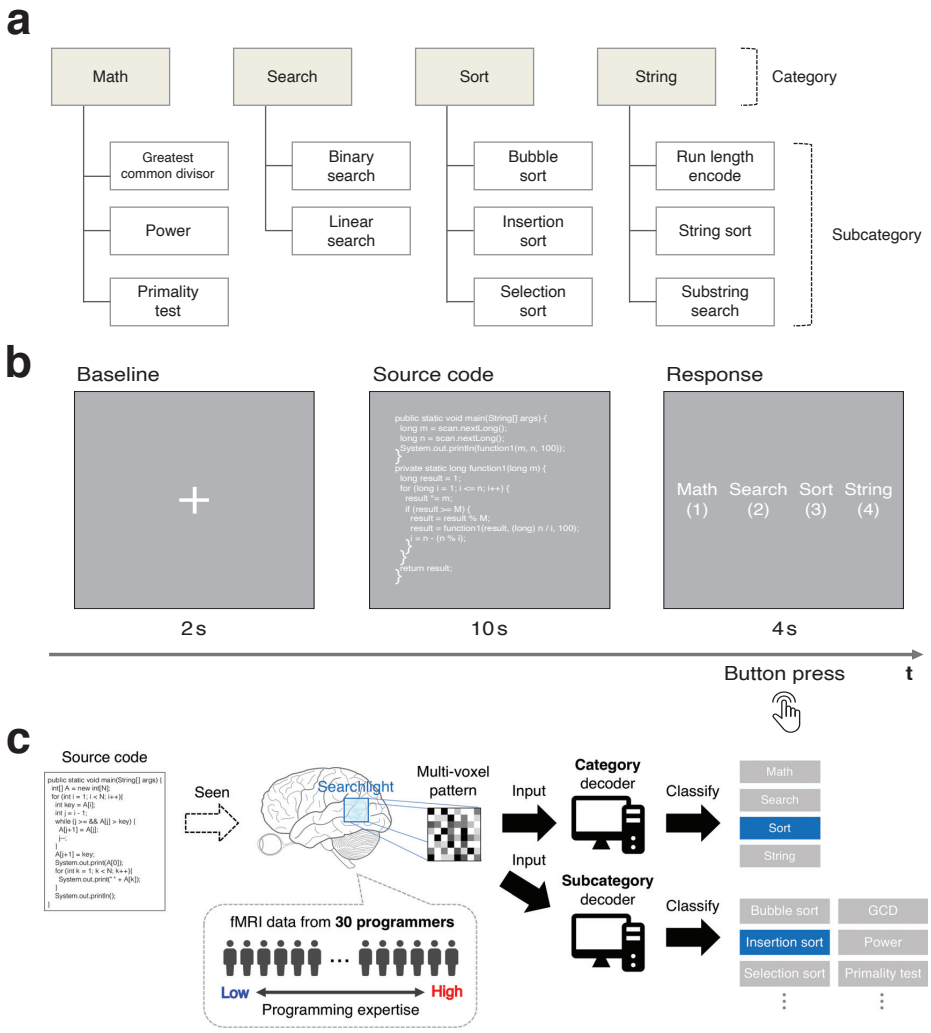


Figure 1: **Experimental design.** (a) Hierarchy of categories used in this study. Category and Subcategory represent abstract functionality and concrete algorithms, respectively, based on two popular textbooks of programming. Every code snippet used in this study belonged to one subcategory class and 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. (c) Overview of the decoding framework. MRI data was collected from 30 subjects with different levels of programming expertise while they performed the program categorization task. Whole-brain searchlight analysis (Kriegeskorte et al., 2006) was employed 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.

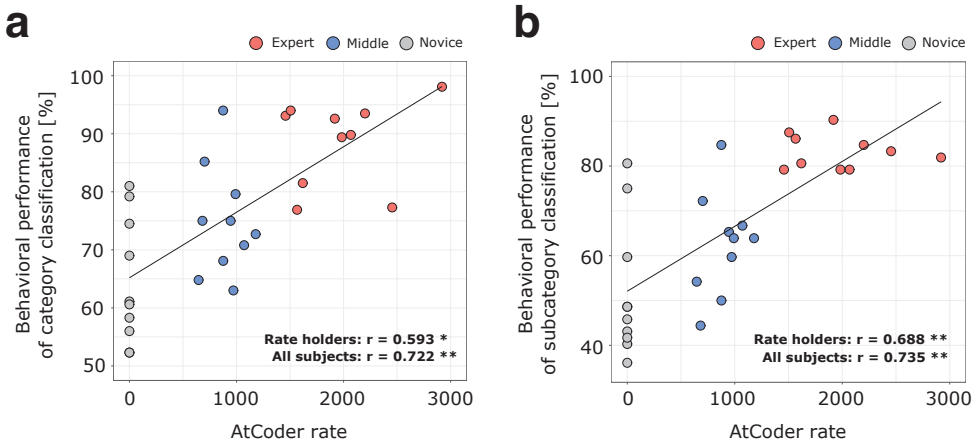


Figure 2: **Correlations between behavioral performance and programming expertise indicator.** (a) Scatter plot of behavioral performances of category classifications against the values of adopted expertise reference (i.e. AtCoder rate). (b) Scatter plot of behavioral performances of subcategory classifications against the values of the same expertise reference. Each dot represents an individual subject. Significance of the correlation coefficients (r) was denoted as *, $p < 0.05$ and **, $p < 0.005$. The solid lines indicate a fitted regression line estimated from all subject data.

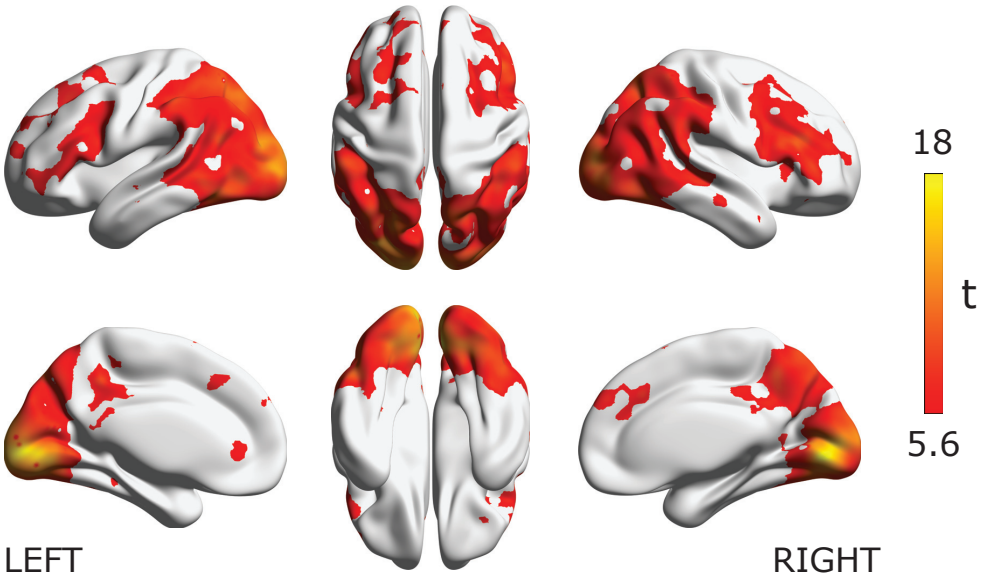


Figure 3: **Decoding accuracy for functional category of source code.** Significant searchlight locations estimated from all subject data (N = 30). Heat colored voxels denote the centers of searchlights with significant decoding accuracy (voxel-level $p < 0.05$, FWE corrected). See Extended Data Fig. 3-1 for the distribution of voxel-level peak decoding accuracies. The brain surface visualizations were performed using BrainNet viewer, version 1.61 (Xia et al., 2013).

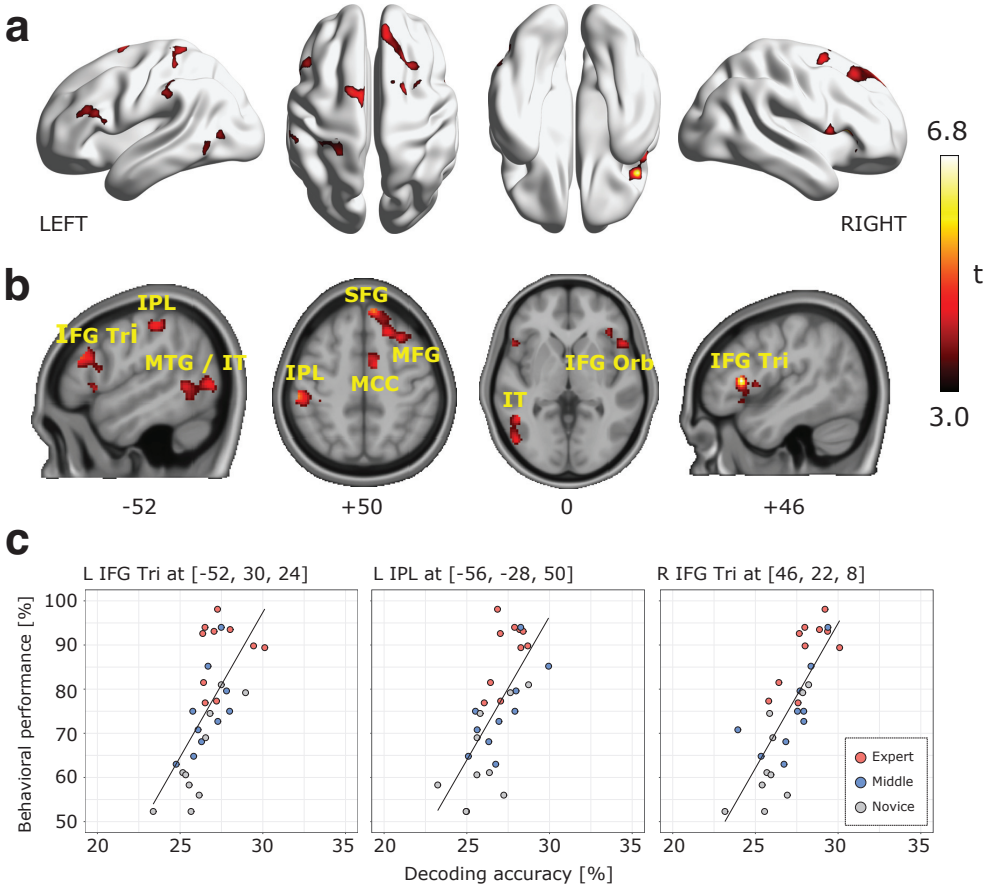


Figure 4: **Searchlight-based correlation analysis between behavioral performances and decoding accuracies.** (a) Locations of searchlight showing significant correlations. Significance was determined by a threshold of voxel-level $p < 0.001$ and cluster-level $p < 0.05$, FWE corrected for the whole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

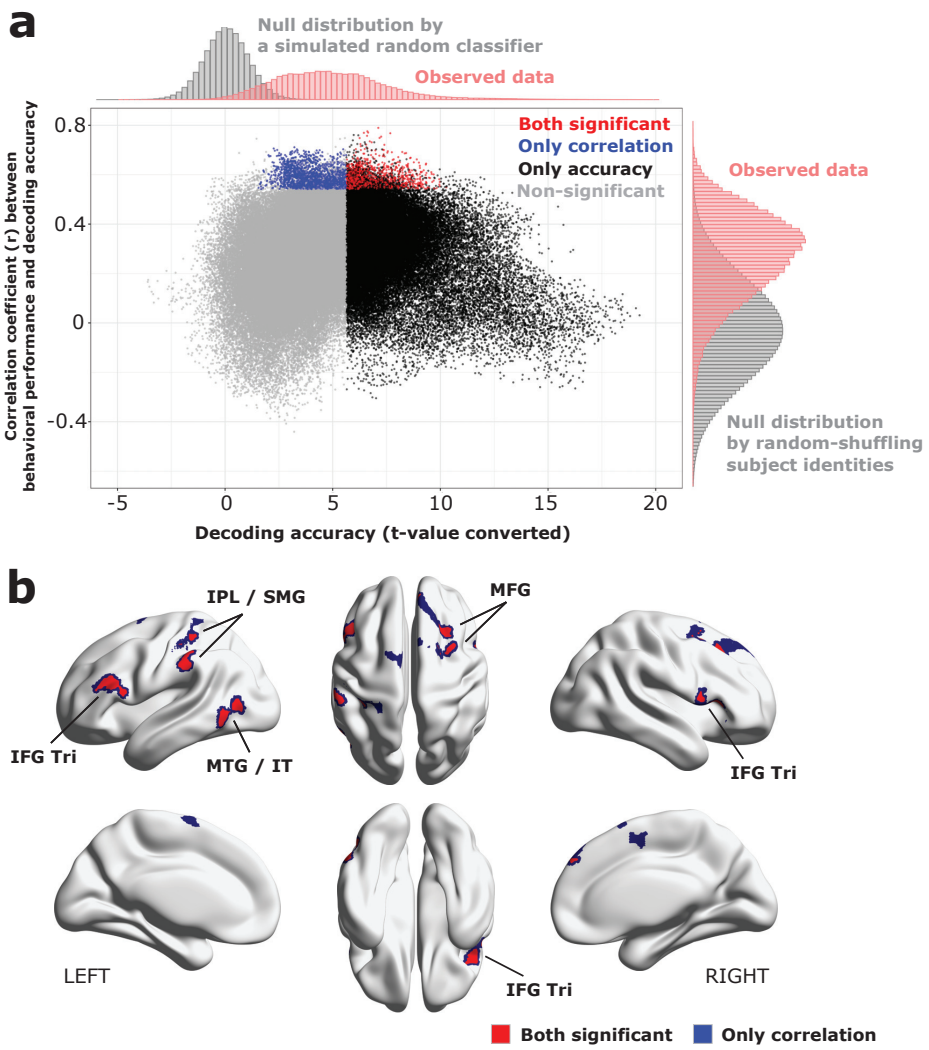


Figure 5: **Identifying searchlight centers that showed both significant decoding accuracy and significant correlation to individual behavioral performances.** (a) Scatter plot of searchlight results. X-axis shows t-values calculated from all subjects' decoding accuracies on each searchlight locations. 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.

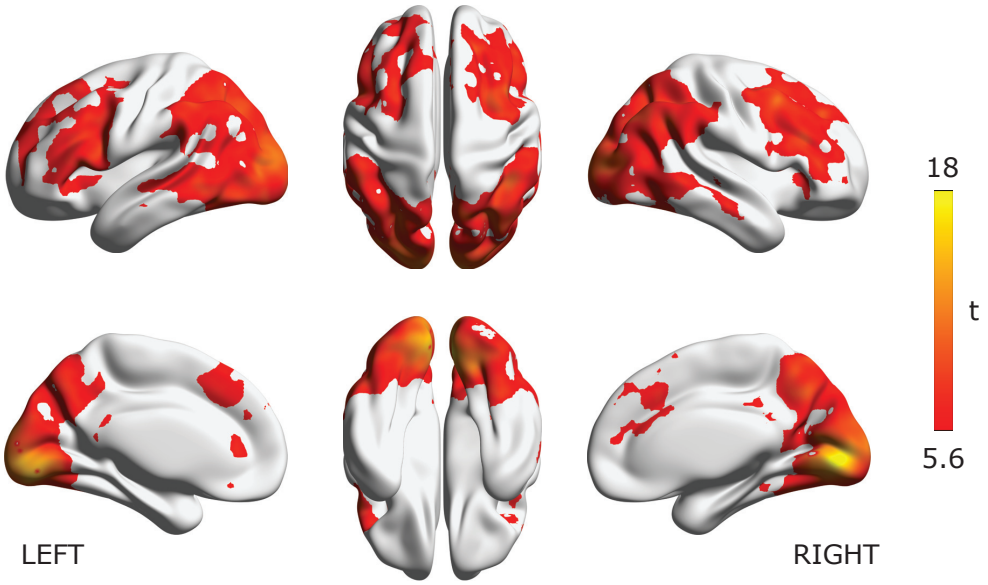


Figure 6: **Decoding accuracy for subcategory of source code.** Searchlight locations showing significant subcategory decoding accuracy than chance estimated from all subject data (N = 30). Heat colored voxels denote the centers of searchlights with significant subcategory decoding accuracy (voxel-level $p < 0.05$, FWE corrected). See Extended Data Fig. 6-1 for the distribution of voxel-level peak subcategory decoding accuracies.

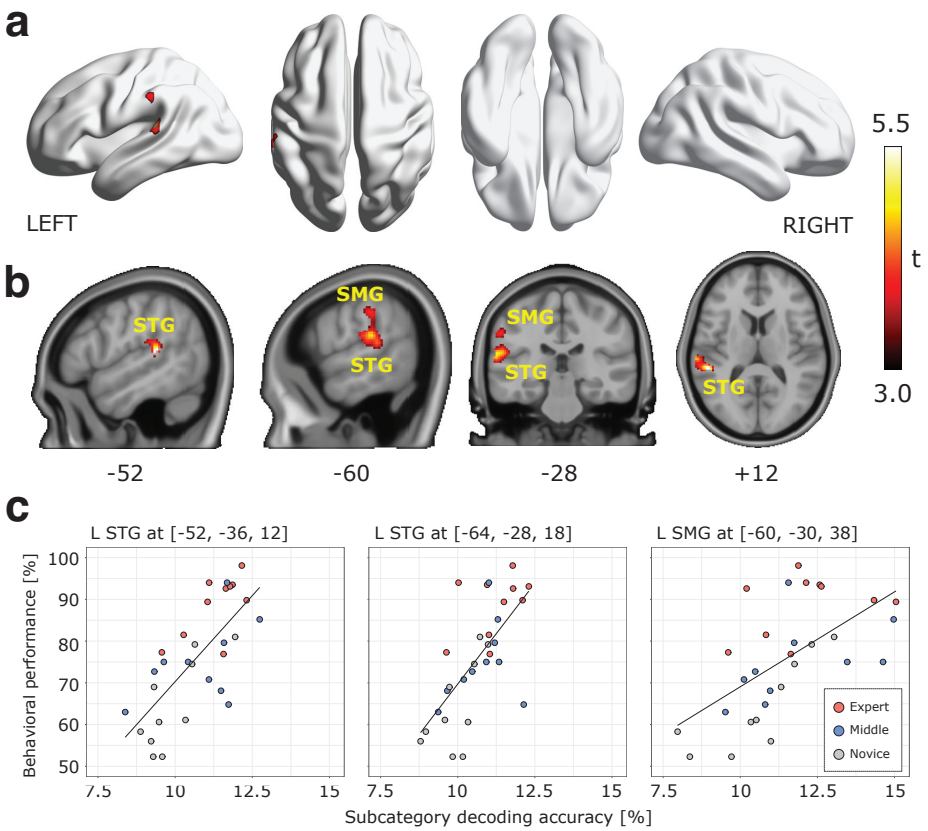


Figure 7: **Searchlight-based correlation analysis between behavioral performances and subcategory decoding accuracies.** (a) Locations of searchlight showing significant correlations. Significance was determined by a threshold of voxel-level $p < 0.001$ and cluster-level $p < 0.05$, FWE corrected for the whole brain. (b) Slice-wise visualizations of the significant clusters. (c) Correlation between behavioral performance and decoding accuracy. Each dot represents an individual subject data. 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.

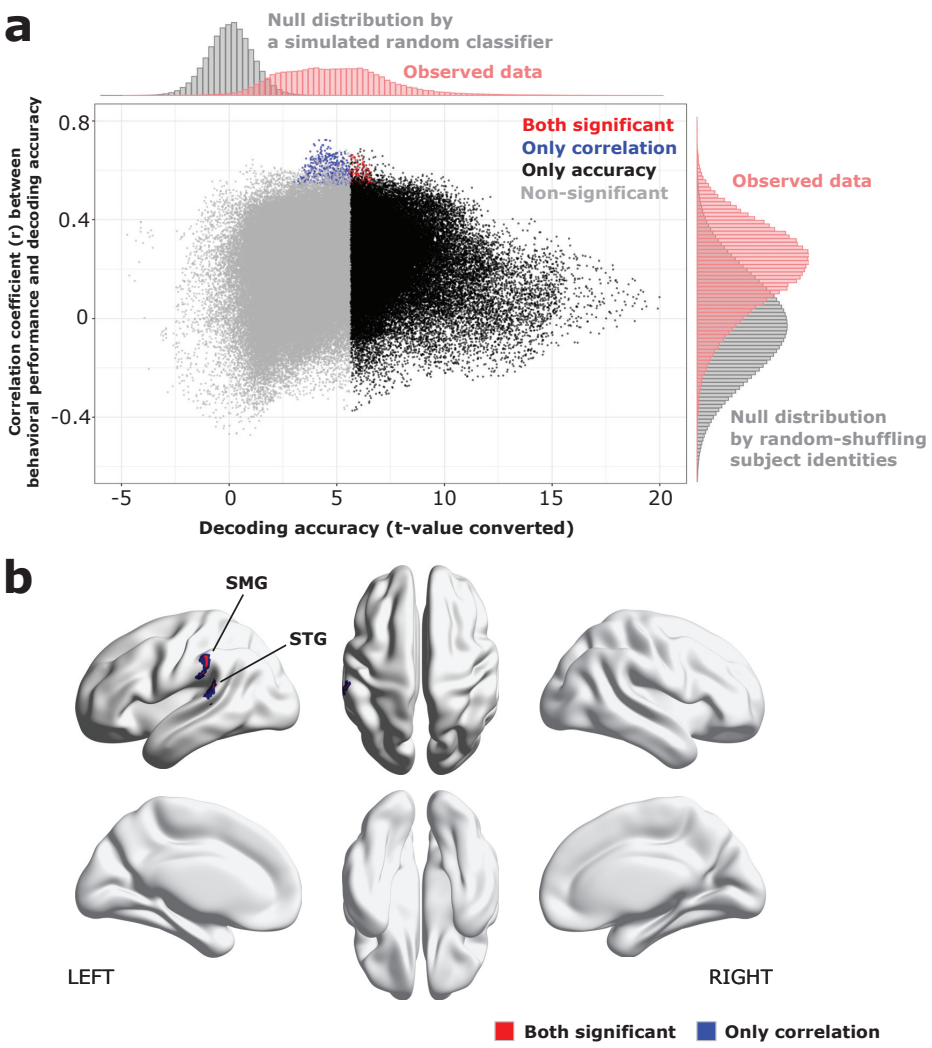
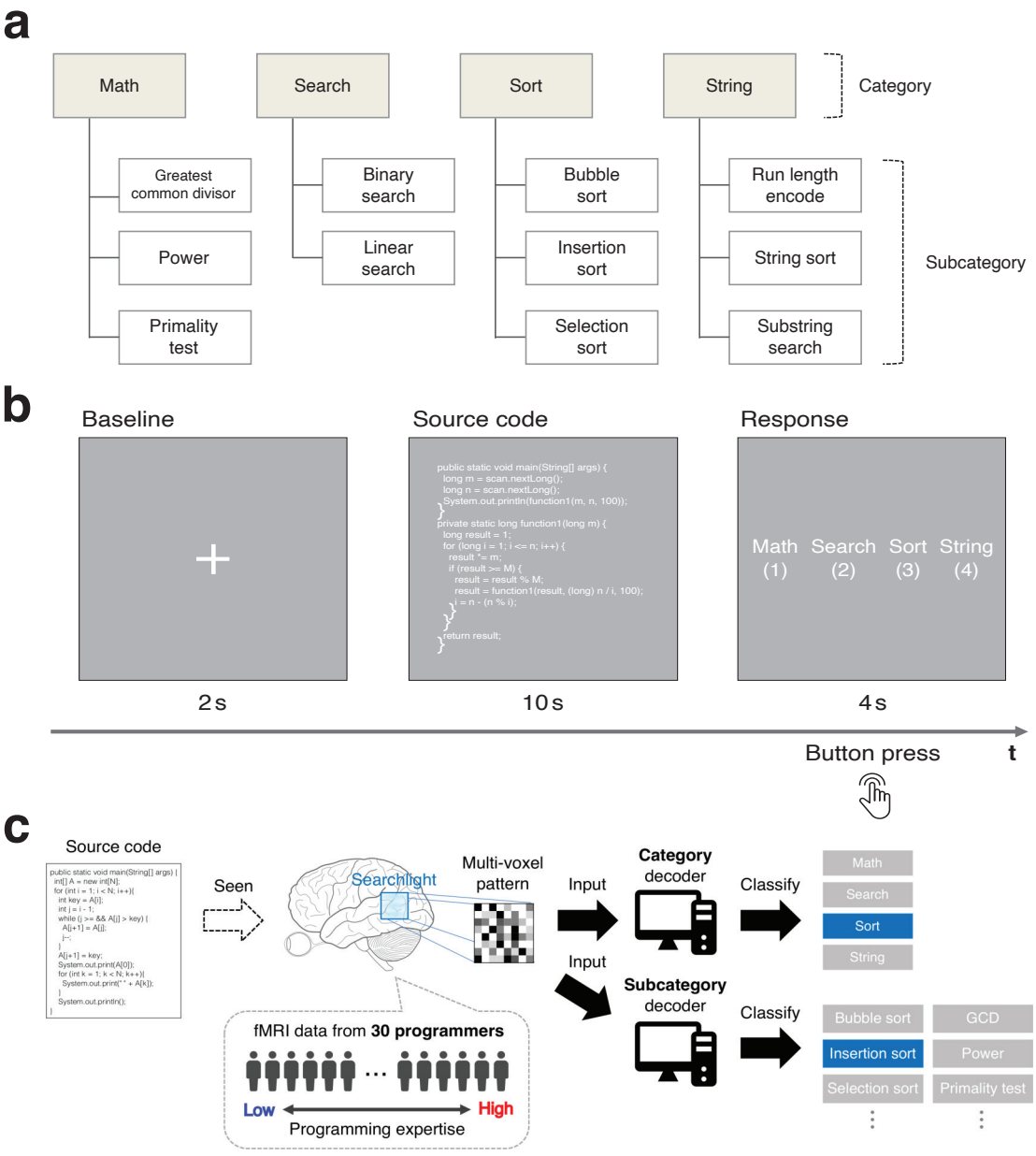
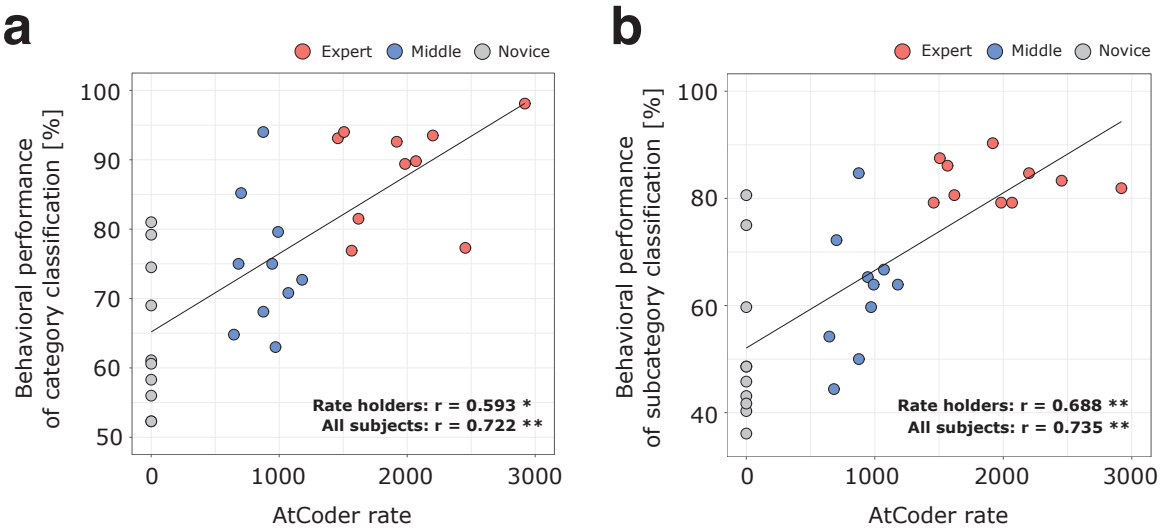
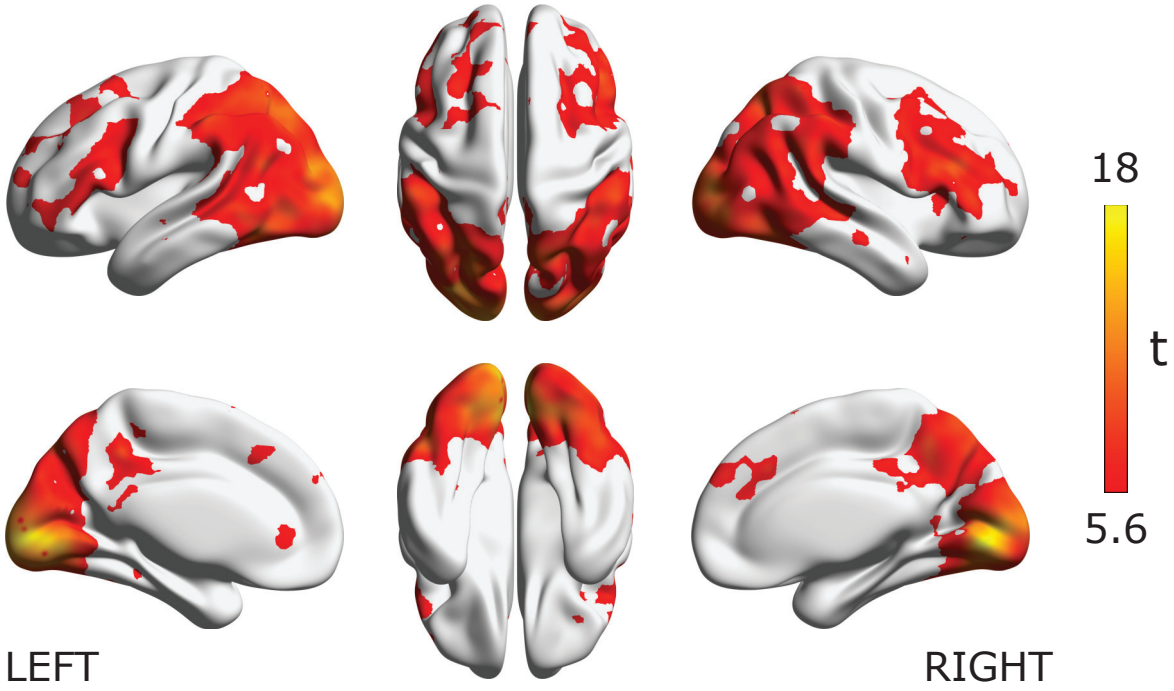
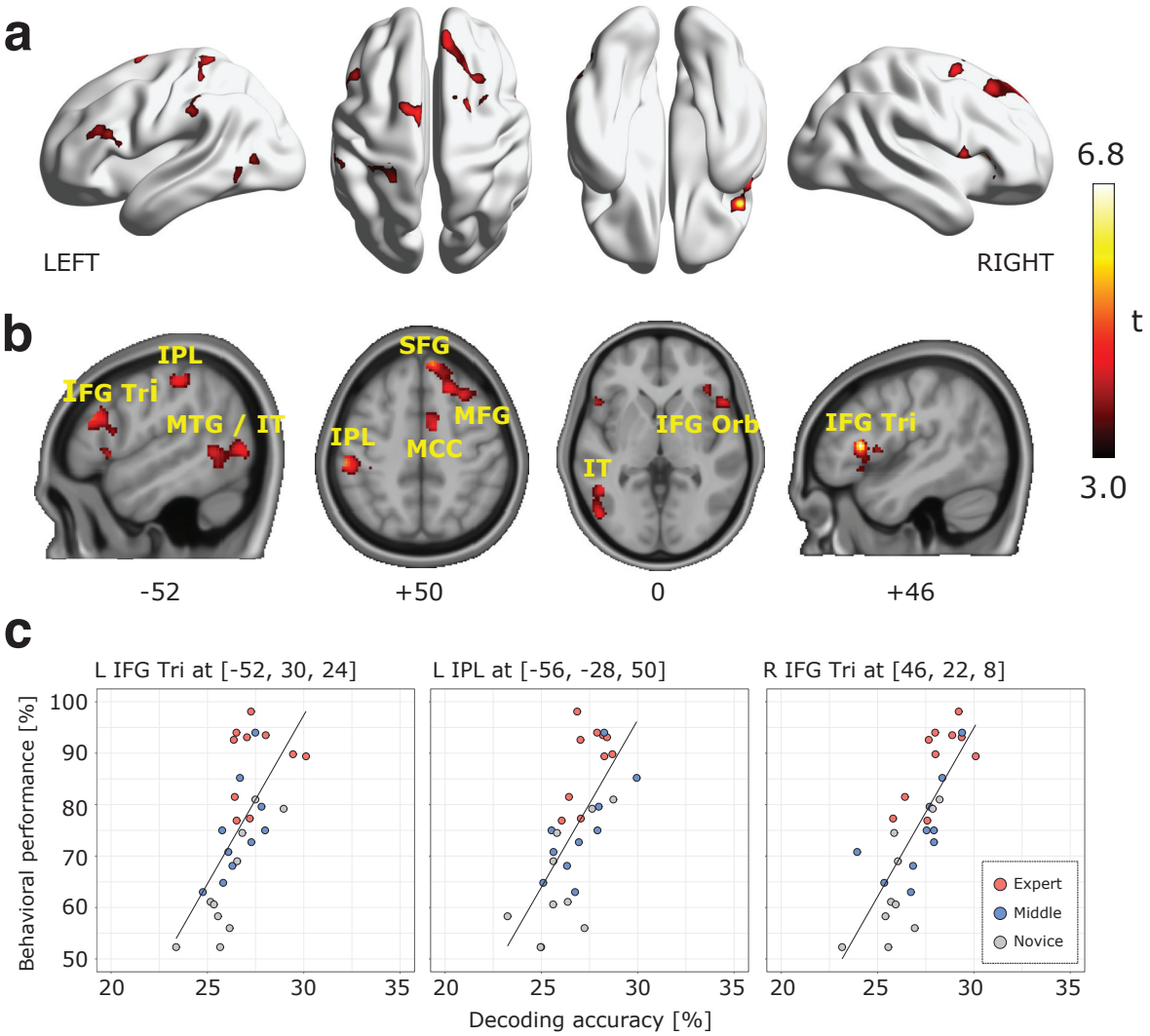


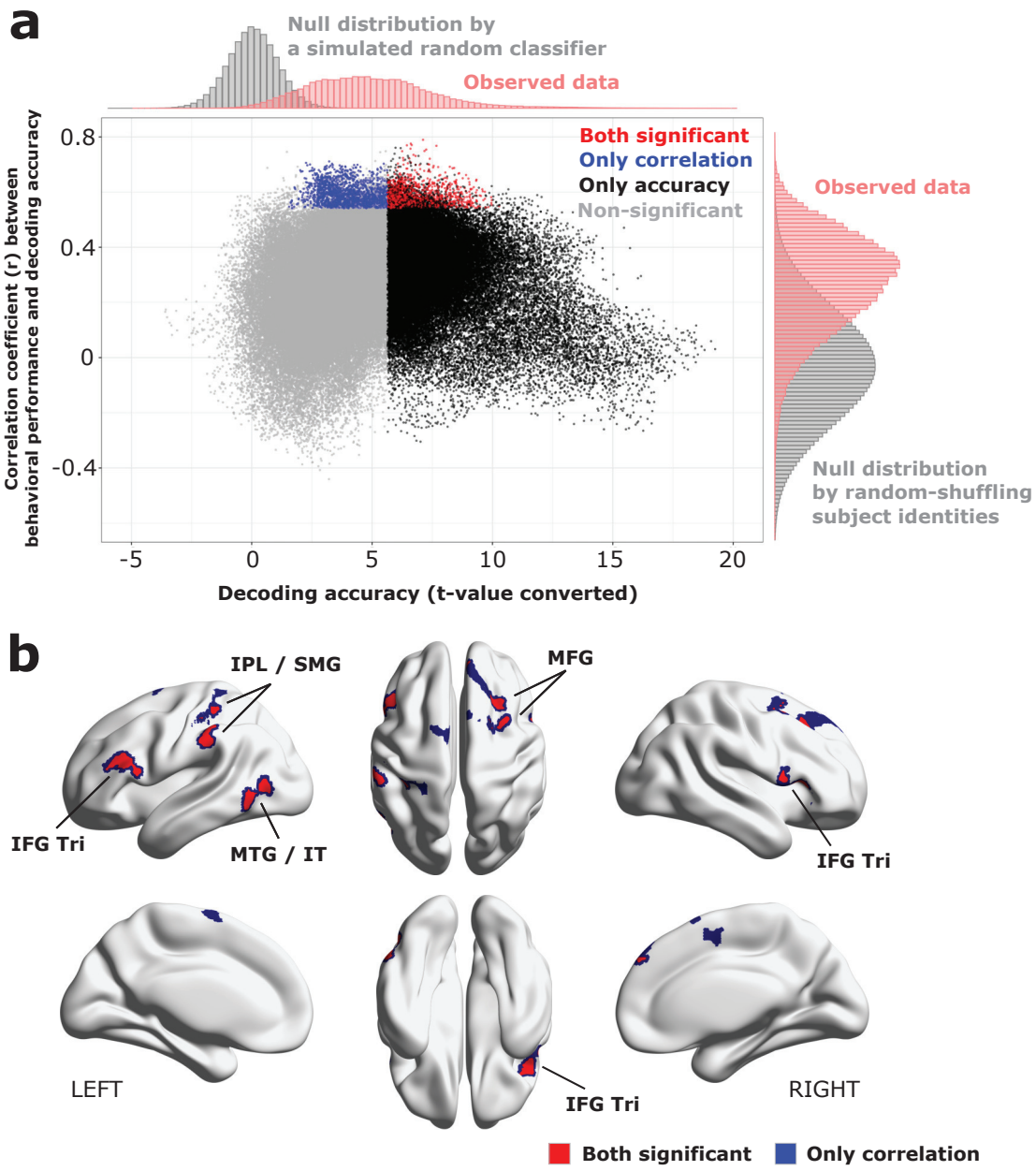
Figure 8: **Identifying searchlight centers that showed both significant subcategory decoding accuracy and significant correlation to individual behavioral performances.** (a) Scatter plot of searchlight results. X-axis shows t-values calculated from all subjects' decoding accuracies on each searchlight 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 subcategory 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.

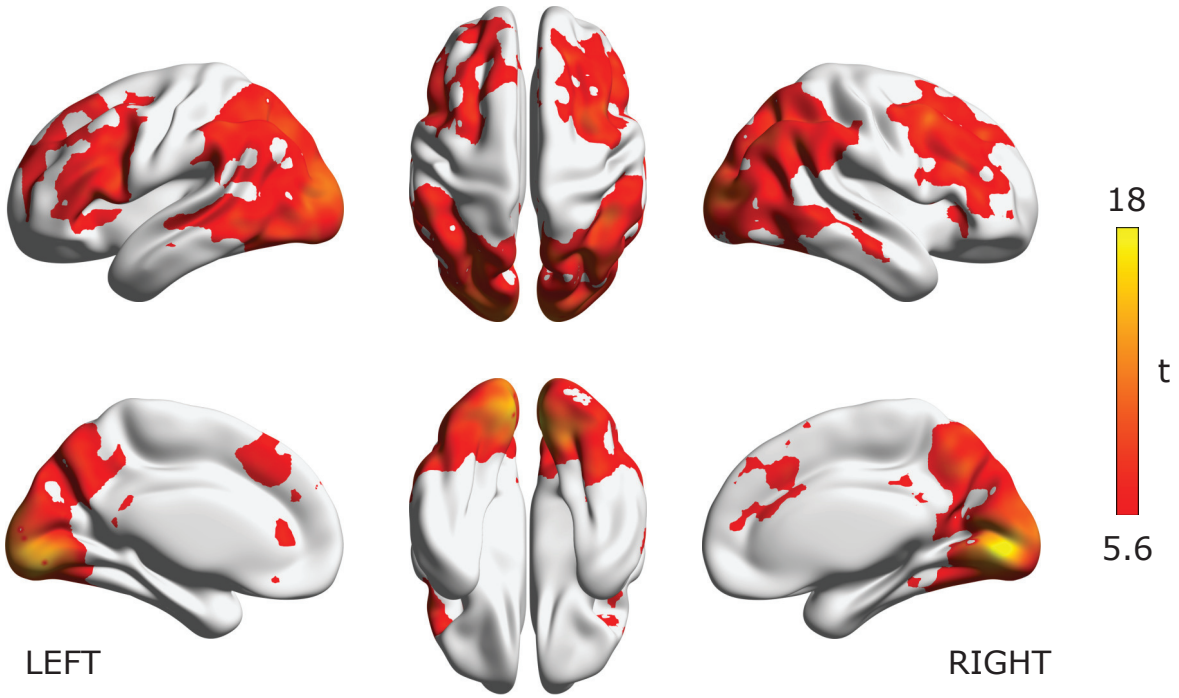


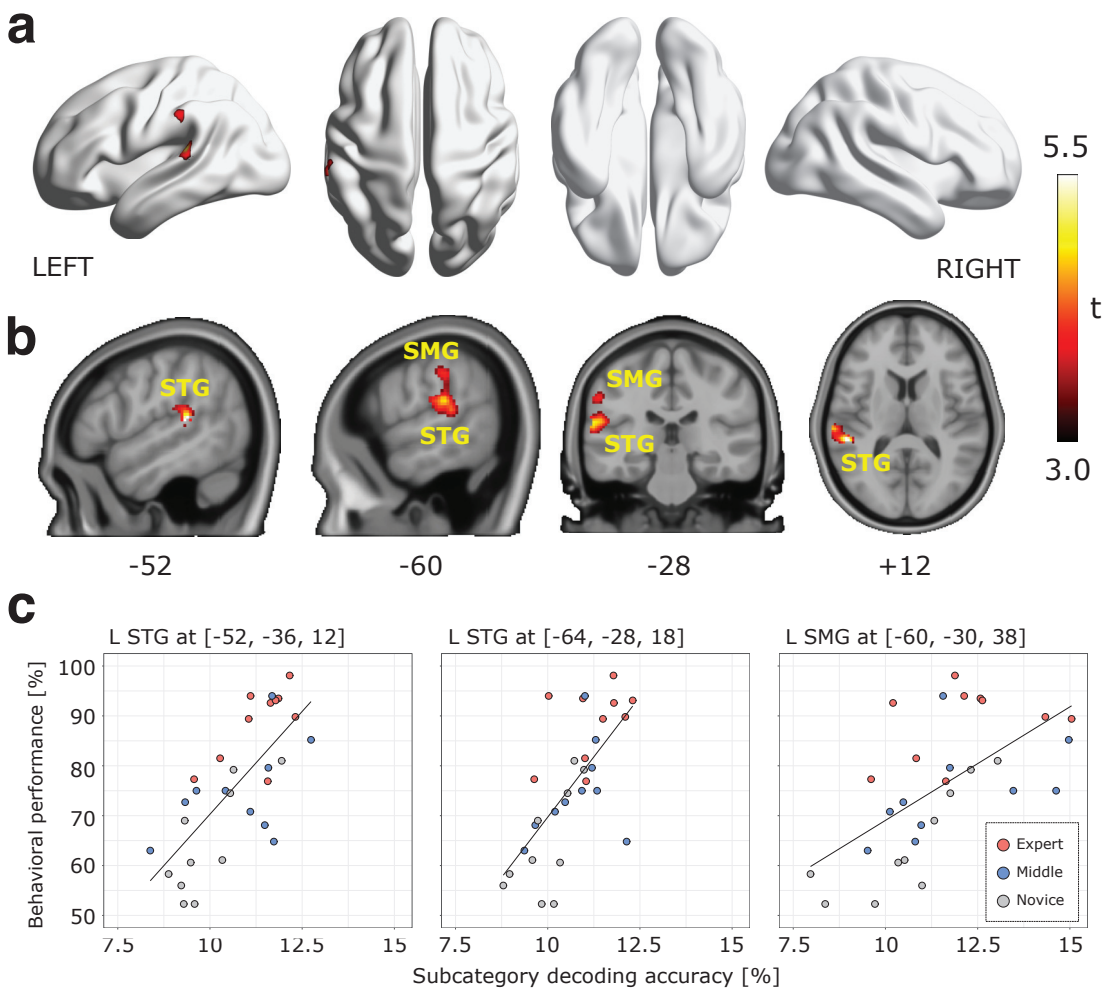












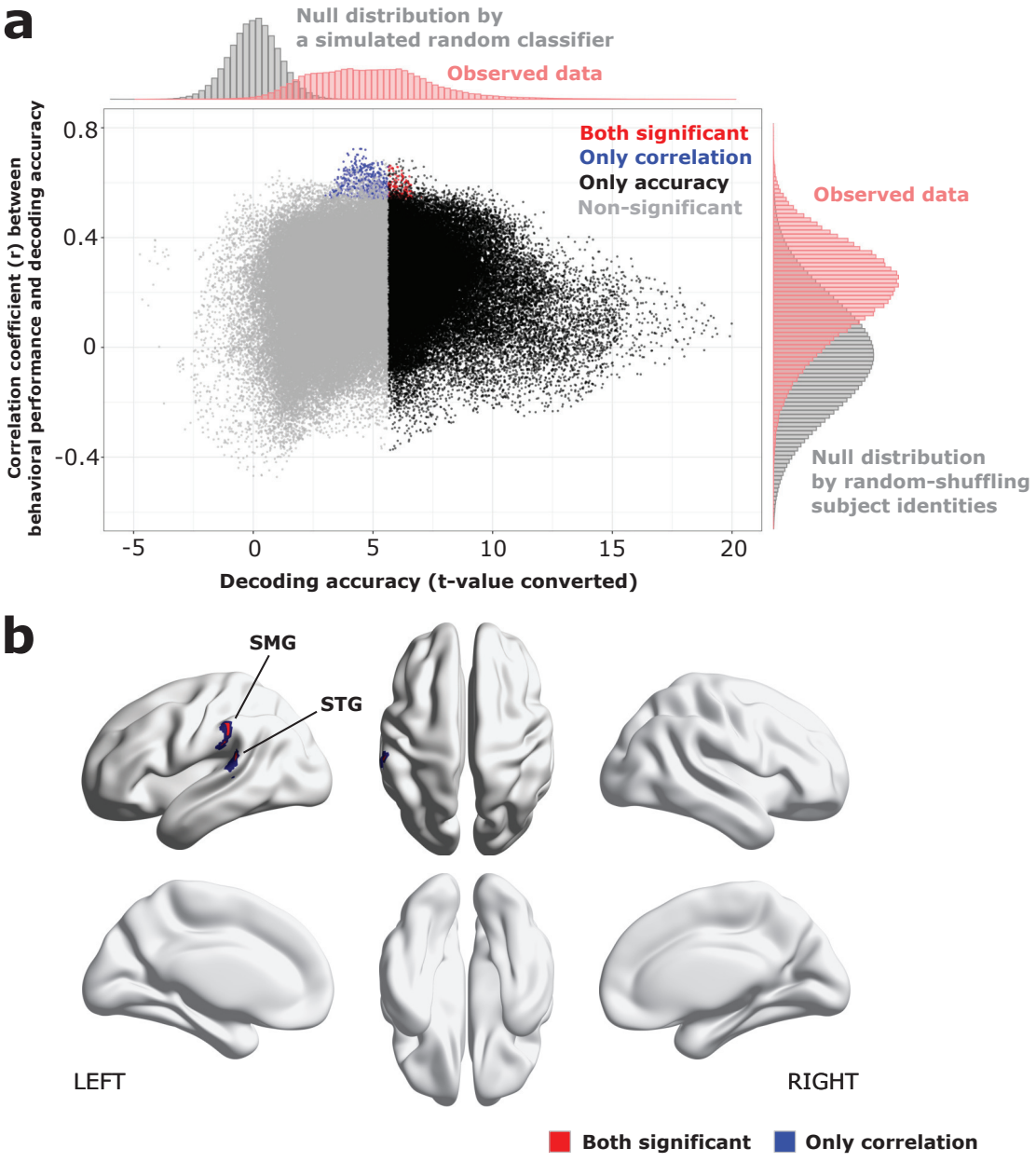


Table 1: **Demographic information of recruited subjects.** Numerics from 4th (Age) to last columns denote 'MEAN ±SD'. Abbreviations: PE, programming experience; JE, Java experience; CPE, competitive programming experience. Significant differences were observed between PE of Expert - Novice, Middle - Novice; CPE of Expert - Middle (two-sample t-test, $p < 0.05$ FDR-corrected).

	N	Sex (M/F)	Age	AtCoder rate	PE (year)	JE (year)	CPE (year)
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

Table 2: **Clusters showing significant correlations between behavioral performance and category decoding accuracy** (voxel-level $p < 0.001$ and cluster-level $p < 0.05$, FWE-corrected). Region names were identified using Automated anatomical labelling atlas 2 (Rolls et al., 2015).

Region name	MNI coordinates			Peak corr. *(r)*	T-value	Cluster extent
	X	Y	Z			
R IFG (p. Triangularis)	46	22	8	0.789	6.81	369
L Posterior-Medial Frontal	-12	0	66	0.711	5.36	298
R Superior Medial Gyrus	6	52	42	0.699	5.17	587
L Inferior Parietal Lobule	-56	-28	50	0.698	5.16	649
R Superior Frontal Gyrus	24	4	60	0.675	4.84	428
L IFG (p. Triangularis)	-52	30	24	0.671	4.79	346
L Inferior Temporal Gyrus	-50	-54	0	0.635	4.35	347