The Quest for the FFA and Where It Led

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.
The Quest for the FFA and Where It Led

Nancy Kanwisher
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

This article tells the story behind our first paper on the fusiform face area (FFA): how we chose the question, developed the methods, and followed the data to find the FFA and subsequently many other functionally specialized cortical regions. The paper’s impact had less to do with the particular findings in the paper itself and more to do with the method that it promoted and the picture of the human mind and brain that it led to. The use of a functional localizer to define a candidate region in each subject individually enabled us not just to make pictures of brain activation, but also to ask principled, hypothesis-driven questions about a thing in nature. This method enabled stronger and more extensive tests of the function of each cortical region than had been possible before in humans and, as a result, has produced a large body of evidence that the human cortex contains numerous regions that are specifically engaged in particular mental processes. The growing inventory of cortical regions with distinctive and often very specific functions can be seen as an initial sketch of the basic components of the human mind. This sketch also serves as a roadmap into the vast and exciting new landscape of questions about the computations, structural connections, time course, development, plasticity, and evolution of each of these regions, as well as the hardest question of all: how do these regions work together to produce human intelligence?

Introduction

In November of 1991, some of the first fMRI images of neural activity in human visual cortex were published on the cover of Science (Belliveau et al., 1991). For me, a psychologist studying visual perception, these images changed everything. Compared with the previous brain-imaging method of positron emission tomography (PET), fMRI images were sharp and crisp, they required no ionizing radiation to the participant, and they could be made every few seconds. Now, scientists could actually watch activity in the normal human brain change over time as it sees, thinks, and remembers. I moved to Boston to try to finagle my way onto (and into) the fMRI scanners at Massachusetts General Hospital (MGH) in Charlestown, then the epicenter of the budding new field.

In the spring of 1995, I finally got my wish: a regular slot on the fMRI scanner at MGH. It was, quite literally, the opportunity of a lifetime. I recruited the two smartest and most committed people I knew to join my team: then-undergraduate Josh McDermott and then-postdoc Marvin Chun. We started off looking for brain regions engaged in visual shape perception. But, after a few months without much success, I became worried. Scan time was expensive, I did not have a grant, and I was at risk of losing my scanner access if my team did not score a big result fast. I figured there was one response that almost had to be lurking somewhere in visual cortex. Extensive evidence from behavior, neurophysiology, and studies of patients with brain damage had already suggested that special brain machinery existed for the perception of faces somewhere in the back of the right hemisphere (Kanwisher and Yovel, 2009). Further, brain-imaging work from both PET and fMRI had already found strong activations when people looked at faces (Sergent et al., 1992; Haxby et al., 1994; Pupe et al., 1996). However, this condition had been compared only to very different control conditions such as viewing letter strings or reporting the locations of visual blobs. The question of whether that region was specifically involved in the perception of faces, rather than more generally engaged in visual shape perception, remained unanswered. I had never worked on face perception because I considered it to be a special case, less important than the general case of object perception. But I needed to stop messing around and discover something, so I cultivated an interest in faces. To paraphrase Stephen Stills, if you can’t answer the question you love, love the question you can.

An initial scan with me as the subject found a promising blob on the bottom of my right hemisphere. Most thrillingly, you could see in the raw time course of the fMRI response of individual voxels that the signal was higher during the periods when I was looking at faces than the periods when I was looking at objects. Still, a single result like that could have been a fluke. So...
Marvin and Josh scanned me again. And again. And again. To our delight, the trusty little blob showed up in exactly the same place every time.

Today, our field faces a replication crisis, with widespread concerns that a substantial proportion of our published findings might be spurious (Szucs and Ioannidis, 2016). I think this problem will be solved not with fancier math, but simply by developing a stronger tradition of replicating our own results before publishing them (especially when those results are surprising).

Replicating results is not glamorous, and falls far short of understanding. But replicating your results is its own special kind of rush. The ability to replicate your phenomenon empowers you to embark on the grander quest of trying to understand it. Replication reveals your own control over the little speck of the universe that you are trying to understand. You are making it come and go at will. You are playing peek-a-boo with nature.

Of course, it is important not just to replicate a phenomenon in the very narrow condition in which you originally found it, but also to generalize it—across people, stimuli, tasks, and methods. When we scanned other people, we found a rather complex pattern, with most subjects showing several different blobs that responded more to faces than objects. What, if anything, was shared across subjects? We did not know about the standard method for answering this question, which was a group analysis. That was lucky: if we had performed a group analysis on those data, we probably would not have found the fusiform face area (FFA) because its location varies too much from one individual to the next.

Instead, we invented our own low-tech version of a group analysis: we taped a paper printout of each subject’s activation map on the wall along the long hall outside my laboratory and we walked back and forth staring at the activation maps, trying to glean the common pattern shared across subjects. It became clear that a few blobs appeared in similar vicinities across subjects, but the most consistent blob was the one on the bottom of the right hemisphere, just above the cerebellum, about an inch in from the skull. We decided to focus on that one first.

Our next problem was that it was not obvious what statistics to run. The activation maps showed that some voxels responded significantly more to faces than objects with impressive \( p \) levels, but these numbers were not corrected for the tens of thousands of statistical tests conducted (one per voxel). On the other hand, a strict Bonferroni correction was clearly too conservative because the voxels were not independent of each other. Software existed that would supposedly do a more appropriate correction, but I didn’t understand it so I didn’t want to use it.

Instead, I decided to do something very simple that I did understand: I split the data in half, using the even runs to find the apparently face-selective blob in each subject and then extracting from the odd runs the average response across the voxels in that blob from the face and object conditions. Now I could run simple \( t \) tests or ANOVAs across subjects on the resulting response magnitudes for faces and objects. No correction for multiple statistical comparisons was necessary because I was running a single statistical test on a single “functional region of interest” or “fROI.” We could also apply the same method to answer new questions: We could use one set of data (from a “localizer” scan with faces and objects) to find the region in each subject and another set of data to measure the response of this region in new conditions of interest. We could do cognitive psychology on a little patch of the brain.

We were not the first to use the fROI method; Roger Tootell and others had been doing something similar, mapping V1 (Sereno et al., 1995) and the visual motion area MT (Tootell et al., 1995) and then separately measuring fMRI responses in those regions in independent data. And, in some sense, neurophysiologists had been using this method for decades: you would obviously first figure out which area your electrode was in before characterizing the single-unit responses measured in that location. It was just common sense.

I still do not understand the resistance to the use of fROIs (Friston et al., 2006), which offer myriad advantages. fROIs enable you to not just make pictures of brain activation, but also to ask principled, hypothesis-driven questions about a thing in nature. In this way, fROIs make possible a cumulative research enterprise in which findings across laboratories can build upon each other because they are studying the same thing (Fedorenko and Kanwisher, 2009) and that thing has systematic, replicable properties, much as one expects for anatomical ROIs such as the amygdala and hippocampus. Standardized brain coordinate systems (e.g., MNI or Talairach coordinates) do not accomplish this goal effectively because functional regions such as the FFA are not well aligned across subjects in standard coordinates. However, a functional localizer allows you to pick out the FFA in each person individually and turn this region into an object of study (including the study of its representations, connectivity, and development). Further, fROIs demonstrate with every use that at least some findings in fMRI are highly replicable across subjects and laboratories. Finally, fROIs enable researchers to avoid three of the most common errors in fMRI data analysis: double dipping (Kriegeskorte et al., 2009), hidden degrees of freedom (Simmons et al., 2011), and invalid methods for correcting \( p \) levels for multiple spatial hypotheses (Eklund et al., 2016).

We could now proceed with the standard scientific method, trying to refute our hypothesis that this region was involved selectively in the perception of faces. Might the region respond, not just to faces, but to any human body part? Or to anything that subjects attend to? Or to any stimulus with the same luminance or contour length or curvature? In subsequent experiments, we localized the candidate face region in each subject individually, measured the magnitude of response in that region to new stimuli testing each of these alternative hypotheses, and ruled out each one.

We quickly wrote a short paper on the work, which was just as quickly rejected from both Science and Nature. We then published the work in the Journal of Neuroscience, which probably increased its impact because we had the room to explain the method in more detail. We weren’t the only ones doing brain-imaging experiments on faces; the same year, Greg McCarthy and colleagues published a study similar to ours (McCarthy et al., 1997), also finding a selective response to faces in the fusiform gyrus. Perhaps the biggest contribution of our paper was its demonstration of a set of methods that enabled you to identify a candidate region, formulate a hypothesis about it, and then test that hypothesis rigorously with multiple repeated tests applied to that same region of the brain.

What was most exciting to me about our work was that it seemed to address directly a major and long-standing theoretical question in cognitive psychology: the degree to which mental architecture is “domain specific,” that is, specialized for particular kinds of information such as faces or places or language. This question had been debated heatedly in our field for nearly 200 years (Finger, 2001) and now here was a little piece of the brain that seemed to do just one thing: perceive faces. This finding fit the broader idea that the mind is not a general purpose device, but is instead composed of a set of distinct components, some
of them highly specialized for solving a very specific problem (Fodor, 1983).

Because the FFA work stepped straight into the middle of this centuries-old debate about domain specificity in the brain, it quickly drew fire from many directions. One argument was that the FFA was not specialized for faces per se, but for the processing of any visual stimulus for which an individual had gained substantial expertise (Gauthier et al., 2000). Although some studies reported higher FFA responses to objects of expertise than control objects (Gauthier et al., 2000; Xu, 2005), consistent with this hypothesis, these effects were small and many other studies failed to replicate them (Grill-Spector et al., 2004; Op de Beeck et al., 2006; Yue et al., 2006). Further, in all studies that have looked, expertise effects are not restricted to the FFA, but extend to multiple other brain regions (Gauthier et al., 2000; Harel et al., 2010; McGugin et al., 2012; Harel et al., 2014), as expected if these effects simply reflect greater attentional engagement by objects of expertise (Harel et al., 2010). Thus, there is no replicable evidence for a special linkage between the FFA (or face selectivity in general) and expertise.

A more serious challenge to the specificity of the FFA came from Jim Haxby, who made the important point that we should care not just about which kinds of stimuli most strongly drive a region, but what information is represented in each region (Haxby et al., 2001). These two things need not be the same, he pointed out, because the pattern of response across voxels within the FFA might be systematically different during viewing of, say, cars versus chairs even if the mean response to the two categories is the same. Indeed, Haxby and others have shown that, by this measure (known as multiple voxel pattern analysis, or MVPA), the FFA does in fact hold information about nonface objects. I consider this the most important current challenge to the specificity of the FFA for faces (and of the other regions discussed below for their preferred categories or functions).

However, what we really want to understand is neither the mean response of the region nor the information content of the neural response, but the causal role of that neural response in behavior. After all, even if this patch of brain were crafted by evolution all and only for representing the difference between one face and another, and even if this is all it was ever used for, it might still produce a different pattern of response to cars versus chairs. This ambiguity is a central problem with MVPA, which shows only the information that we scientists can fish out of the response of a given patch of brain, not the information that the rest of the brain is reading out of that patch of brain (Williams et al., 2007). The only way to determine the causal role of that brain region in behavior is to intervene on it. So far, results from intervention studies (including brain damage, transcranial magnetic stimulation, and electrical stimulation) indicate that category-selective regions of the brain are primarily or exclusively causally engaged in representing their preferred stimulus categories (Pitcher et al., 2009). However, new and more precise methods for causal intervention on neural representations (Afraz et al., 2015) are being developed that should provide stronger tests of these ideas in the next few years.

The localize-and-test fROI method that we developed in the FFA paper proved useful for identifying and characterizing a number of other functionally specific regions of cortex. Following up on the mysterious “negative activation” from the FFA paper, a higher response to objects than faces, Russell Epstein and I found the parahippocampal place area (PPA), which responds selectively to images of places (Epstein and Kanwisher, 1998). A few years later, Paul Downing and I found the extrastriate body area (EBA), which responds selectively to bodies (Downing et al., 2001). However, in a systematic test that Paul ran on 20 different kinds of object categories, including tools, flowers, spiders, and snakes, we did not find other highly specialized regions.

Evidently, we do not have specialized brain regions for every category of object that we can perceive. Why do some categories get their own private patch of real estate in the brain while others do not? Is it just visual categories that have longstanding evolutionary significance that get their own region? Evidently not. A tiny region near the face area, but in the left hemisphere, responds selectively to visually presented letter strings, but only if you know how to read (Saygin et al., 2016) and only for an orthography you know (Baker et al., 2007). The existence of this “visual word form area” shows that at least one region of the cortex has a strong selectivity that cannot be innate (Polk and Farah, 1998), but is instead based on the experience of that individual. Whether all cortical selectivities develop through experience in this way or whether some are innate remains a fundamental and unanswered question.

Surprisingly basic questions remain unanswered about the information represented in each region, the computations it conducts, and the connectivity between that region and the rest of the brain. Methods exist that are capable of at least approaching some of these questions in humans, but each has substantial limitations. fMRI adaptation has taught us much about what is represented in each region of the brain, but the mechanisms behind adaptation, and thus the interpretation of results of adaptation studies, are open to multiple interpretations. Although widely used and undeniably elegant, MVPA often produces very low decoding accuracies (e.g., 55% correct where chance is 50%). Efforts in my laboratory have shown zero ability to decode identity, race, or gender from the FFA, even with very high-resolution scans (see also Jeong and Xu, 2016), although some other studies have managed such decoding, albeit with relatively low accuracy (Anzellotti, Fairhall, and Caramazza, 2014; Axelrod and Yovel, 2015; Guntupalli et al., 2016). We know that the information is there but, perhaps not surprisingly, we often can’t see it with a method that averages responses across hundreds of thousands of neurons in each voxel (Dubois et al., 2015). Given the weaknesses in the methods available for investigations in humans, I began to despair that fundamental questions about face-selective patches in the brain might just never be answered.

But help was on the way. In 2003, Doris Tsao and Winrich Freiwald discovered face-selective patches of cortex in macaque monkeys using fMRI methods much like those we had used in humans. Now, precise questions about the computations of face-selective patches could finally be answered. Indeed, a few years later, they published an even more exciting result: by directing electrodes into the face patches in monkeys (identified with fMRI), they found that the vast majority of individual neurons in the macaque face patches responded extremely selectively to faces. This result lent powerful support to the earlier work in humans, showing that the selectivity was even stronger than suggested by fMRI, but even more importantly, it enabled Tsao and Freiwald to look directly at the actual neural code for faces, which we can only “see” in drastically blurred form with fMRI. Over the next few years, Tsao and Freiwald and their colleagues swiftly answered many of the fundamental questions that the work on humans had been unable to answer. They showed the progression in face representations across hierarchically organized face patches (Freiwald and Tsao, 2010), the time course of response in each patch, and the precise connectivity of the face system, revealed by tracer injection studies (Grimaldi et al., 2016) and by
Human neuroscience has come far in the last quarter century. Figure 1 (top) shows the approximate state of knowledge of the functional architecture of the human brain since 1990, just before the invention of fMRI. It was not at all obvious then that more functional structure existed, waiting to be discovered. The picture on the top could have remained the whole story, with no other brain regions selectively engaged in specific mental functions at all. Instead, the glorious picture that has emerged from fMRI research (bottom of Fig. 1) shows a large number of functionally specific regions of the cortex, each of which has been widely replicated in many different laboratories. Functional imaging of the brain has begun to reveal, in a very concrete way, the functional organization of the human mind.

This general picture is not universally accepted in the field. Some of the disagreement results, I think, from a simple misunderstanding of the concept of functional specificity, which is often conflated with other ideas. Brain specialization is often assumed to imply innateness, yet, as the case of the visual word form area shows, these concepts are independent. It is an open (and fascinating) question which of the functionally specific regions in the rTPJ has been shown to respond selectively when you think about another person’s thoughts (Saxe and Kanwisher, 2003). The classic brain regions for language, which are strongly activated when you understand the meaning of a sentence, turn out to be inactive when you perform arithmetic, hold information in working memory, exert “cognitive control” or listen to music, showing that, as far as the brain is concerned, language and thought are not the same thing (Fedorenko and Varley, 2016). The discovery that distinct populations of neurons respond selectively to speech and music (Norman-Haignere et al., 2015) shows that music is not simply a byproduct of speech, but rather its own separate thing in the brain. These brain regions, specialized for core components of human cognition, afford a new window into human nature.

Lest it seem that every mental function has its own special brain region or vice versa, that is decidedly not the case. Indeed, just as remarkable as the extreme specificity of the regions described above is the almost disreputable indiscriminateness of another set of brain regions often referred to as “multiple demand” regions, which respond to almost any kind of task demand (i.e., difficulty; Duncan and Owen, 2000; Fedorenko et al., 2013). Conversely, many of the mental functions that we and others have studied engage brain regions that are not specialized for that function alone. For example, our ability to infer intuitively the basis of physical events appears to engage brain regions also known for their role in planning actions (Fischer et al., 2016). Ultimately, these cases in which two apparently different mental functions cohabit the same brain region may prove most informative about the representations and computations that underlie these mental abilities.
are not arbitrary things (or “natural kinds”; Quine, 1969), but rather arbitrary subdivisions of underlying continua (Op de Beeck et al., 2008; Huth et al., 2012). Of course, a complex system such as the brain can be subdivided along multiple dimensions and levels of analysis (Marr, 1982) and there is no single privileged organizing scheme. Instead, the nature and grain of the most useful units in one’s theories depend on the phenomena that one is attempting to explain. I would argue that, for an understanding of the human mind, functionally specific brain regions do in fact carve nature at its joints, capturing structure inherent in both cognitive and neural data. Many functionally specific cortical regions represent not just peaks of broad functional selectivities spanning centimeters of cortex, but relatively sharp spatial discontinuities in functional responses along the cortical surface. For example, the signature selectivity of the FFA and PPA drops to zero within 4 mm outside of the standardly defined border of these regions (Spiridon et al., 2006). Further, growing evidence indicates that each functionally defined cortical region has a distinctive pattern of connectivity to the rest of the brain (Saygin et al., 2011; Osher et al., 2016) and some of these regions may even correspond to cytoarchitectonic divisions of the cortex (Weiner et al., 2014). Therefore, current evidence supports the idea that these regions are distinct in their spatial borders, functional responses, connectivity, causal role in behavior, and perhaps also cytoarchitecture, thus already meeting many of the classical criteria for cortical areas (Felleman and Van Essen, 1991). But the real test of a natural kind is whether it can explain future data. It is an open and exciting empirical question whether the functionally distinct regions of the cortex argued for here will turn out to correspond to discontinuities in other kinds of data such as trajectories of development, patterns of gene expression, and the computational architecture of cognition.

fMRI has opened up a vast landscape of fundamental new questions. What is the time course of processing in each functionally specific cortical region and how do these regions interact with each other online during processing? What is the causal role of each region in cognition and behavior? How fixed are these regions in adulthood and when and how can they reorganize after brain damage? When (Deen et al., 2016) and how (Saygin et al., 2016) does each region arise over development? Perhaps most fundamentally, what is the evolutionary origin of the brain regions that implement distinctively human functions such as language, music, and understanding other minds?

For many of these questions, currently available methods in humans are likely insufficient, although intracranial recording (Allison et al., 1994; Fedorenko et al., 2016) and stimulation (Parvizi et al., 2012) in neurosurgery patients are particularly informative when available. We will have to keep scoping the horizon, looking for opportunities to chip away at these big questions. For whatever measurements we make, it will behoove us to first identify functionally where we are in the brain so that our objects of study are actual things in nature. Indeed, the most important legacy of the FFA may be the establishment of a cumulative research program on the human mind and brain in which the analyses that we conduct are principled and the questions that we ask—and the answers that we receive—are meaningful.

References


