LETTERS
https://doi.org/10.1038/s41550-018-0587-9

nature astronomy

A giant protocluster of galaxies at redshift 5.7

Linhua Jiang[1*], Jin Wu[1,2], Fuyan Bian[3], Yi-Kuan Chiang[4], Luis C. Ho[1,2], Yue Shen[5,6], Zhen-Ya Zheng[7,8,9], John I. Bailey III[10,11], Guillermo A. Blanc[12,13], Jeffrey D. Crane[12], Xiaohui Fan[14], Mario Mateo[10], Edward W. Olszewski[14], Grecco A. Oyarzún[13], Ran Wang[1] and Xue-Bing Wu[1,2]

Galaxy clusters trace the largest structures of the Universe and provide ideal laboratories for studying galaxy evolution and cosmology[1,2]. Clusters with extended X-ray emission have been discovered at redshifts of up to $z \approx 2.5$ (refs [3–7]). Meanwhile, there has been growing interest in hunting for protoclusters, the progenitors of clusters, at higher redshifts[8–14]. It is, however, very challenging to find the largest protoclusters at early times, when they start to assemble. Here, we report a giant protocluster of galaxies at $z \approx 5.7$, when the Universe was only one billion years old. This protocluster occupies a volume of about 35^3 cubic comoving megaparsecs. It is embedded in an even larger overdense region with at least 41 spectroscopically confirmed, luminous Lyα-emitting galaxies (Lyα emitters, or LAEs), including several previously reported LAEs[9]. Its LAE density is 6.6 times the average density at $z \approx 5.7$. It is the only one of its kind in an LAE survey in 4 deg^2 on the sky. Such a large structure is also rarely seen in current cosmological simulations. This protocluster will collapse into a galaxy cluster with a mass of $(3.6 \pm 0.9) \times 10^{15}$ solar masses, comparable to those of the most massive clusters or protoclusters known so far.

According to cosmological simulations, the largest protoclusters of galaxies extend over tens of comoving megaparsecs (cMpc) at $z > 5$ (refs [15,16]). Deep wide-area surveys are needed to find these giant structures at high redshift. We are carrying out a spectroscopic survey of galaxies in 4 deg^2 on the sky, aiming to build a large and homogeneous sample of LAEs at $z \approx 5.7$ and $z \approx 6.5$. We are observing five well-studied fields, including the Subaru XMM-Newton Deep Survey (SXDS) field[17]. These fields have deep optical imaging data in a series of broad and narrow bands, taken by the prime-focus imager Suprime-Cam on the 8.2 m Subaru Telescope. The combined SXDS images in five broad-band filters (B, V, R, i' and z', centred at 442 nm, 545 nm, 650 nm, 763 nm and 903 nm, respectively) and two narrow-band filters (NB816 and NB921, centred at 816 nm and 921 nm, respectively) have enabled us to efficiently select LAE candidates at $z \approx 5.7$ and $z \approx 6.5$ via the Lyα technique[18–21]. From these LAE candidates, we identified a large overdense region at $z \approx 5.7$ in the SXDS. Here, we show that this overdense region contains a giant protocluster (SXDS_gPC) that will grow into a massive galaxy cluster.

We carried out deep spectroscopic observations of SXDS_gPC using the Michigan/Magellan Fiber System (M2FS)[22], a fibre-fed multi-object spectrograph on the 6.5 m Magellan Clay Telescope. M2FS has 256 optical fibres deployed over a circular field of view 29.2 arcmin in diameter. SXDS_gPC and its surrounding environment were covered by one M2FS pointing that included $z \approx 5.7$ LAE candidates brighter than NB816 = 25.8 mag (5σ detection for point sources; magnitudes are on the AB system), as well as a variety of galaxy candidates at other redshifts. We used a pair of red-sensitive gratings with a resolving power of about 2,000. We obtained 7 h of on-source integration (seven 1 h individual exposures) during dark time in November 2015. The combined spectrum reaches a Lyα flux depth of $\sim 1 \times 10^{-17}$ erg s^{-1} cm^{-2} (or a luminosity depth of $\sim 4 \times 10^{42}$ erg s^{-1}, assuming $H_0 = 68$ km s^{-1} Mpc^{-1}, $\Omega_m = 0.3$ and $\Omega_\Lambda = 0.7$, where H_0 is the current value of the Hubble constant, and Ω_m and Ω_Λ are the cosmological density parameters for matter and dark energy, respectively), which ensures reliable identification of LAEs down to at least NB816 = 25.5 mag. See our programme overview paper[23] and Methods for details on the imaging data, target selection and M2FS observations.

The spectroscopic observations have allowed us to remove contaminants from the LAE candidates and measure accurate redshifts for confirmed LAEs. Both are critical to characterize protoclusters. We confirmed 46 luminous LAEs at $z \approx 5.7$ from our spectroscopic data. Example spectra of four LAEs are given in Fig. 1. More details of the 46 LAEs are provided in Supplementary Fig. 1 and Supplementary Table 1. Figure 2 illustrates the SXDS field, the LAE locations and the M2FS pointing, with an area coverage of 660 arcmin^2. The spatial distribution of the 46 LAEs is highly uneven: 41 of them are located in the southwest half of the circular field, including some previously reported LAEs[9]. These 41 LAEs belong to the large overdense region that we identified from the photometric data. In Fig. 2, the projected area of this overdense region is enclosed by the half-circle-like shape outlined in magenta. Its size is roughly 370 arcmin^2, or approximates to 53×41 cMpc^2. Its real size may be larger, because the current size estimate is limited by the coverage of the Subaru imaging data and the M2FS spectroscopic data.

Figure 3 shows the redshift distribution of the LAEs. They span an effective redshift interval of $\Delta z \approx 0.10$, which corresponds to a line-of-sight depth of ~46 cMpc. The redshift distribution implies the existence of two groups with a slight redshift offset. The majority of the LAEs are in a group at $z \approx 5.68$, with $\Delta z \approx 0.075$, or a line-of-sight depth of ~34 cMpc. The remaining LAEs are in a narrower

[1]Kavli Institute for Astronomy and Astrophysics, Peking University, Beijing, China. [2]Department of Astronomy, School of Physics, Peking University, Beijing, China. [3]Research School of Astronomy and Astrophysics, Australian National University, Weston Creek, Australian Capital Territory, Australia. [4]Department of Physics & Astronomy, The Johns Hopkins University, Baltimore, MD, USA. [5]Department of Astronomy, University of Illinois at Urbana-Champaign, Urbana, IL, USA. [6]National Centre for Supercomputing Applications, University of Illinois at Urbana-Champaign, Urbana, IL, USA. [7]CAS Key Laboratory for Research in Galaxies and Cosmology, Shanghai Astronomical Observatory, Shanghai, China. [8]Institute of Astrophysics and Center for Astroengineering, Pontificia Universidad Catolica de Chile, Santiago, Chile. [9]Chinese Academy of Sciences South America Center for Astronomy, Santiago, Chile. [10]Department of Astronomy, University of Michigan, Ann Arbor, MI, USA. [11]Leiden Observatory, Leiden University, Leiden, The Netherlands. [12]Observatories of the Carnegie Institution for Science, Pasadena, CA, USA. [13]Departamento de Astronomía, Universidad de Chile, Santiago, Chile. [14]Steward Observatory, University of Arizona, Tucson, AZ, USA. *e-mail: jiangKIAA@pku.edu.cn

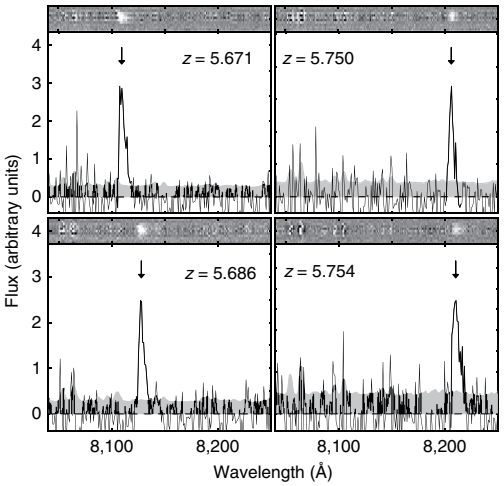

Fig. 1 | Example spectra of four LAEs taken by M2FS. In each case, we show the two-dimensional (upper) and one-dimensional (lower) spectra, with the zero-flux level (dashed line) and 1σ uncertainty region (grey) indicated on the one-dimensional spectrum. The downward arrow points to the position of the Lyα emission line. The spectra of all 46 LAEs are shown in Supplementary Fig. 1.

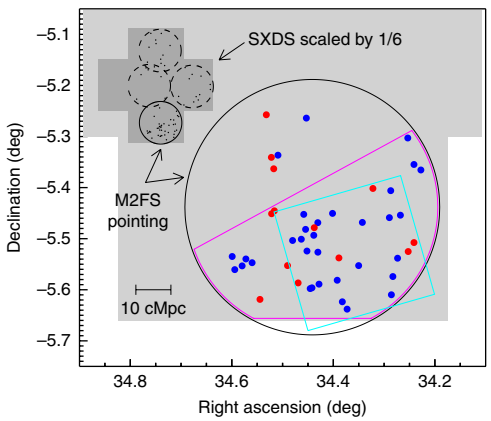

Fig. 2 | Schematic representation of the SXDS_gPC region. The light grey background shows the coverage of the Subaru imaging data, and the black circle indicates the coverage of the M2FS pointing for SXDS_gPC. The whole SXDS field, scaled by 1/6, is overplotted in dark grey in the upper-left corner, where the dashed circles plot another three M2FS pointings observed by our M2FS programme. The black points represent all spectroscopically confirmed LAEs brighter than NB816 25.5 mag, which clearly indicate a large overdense region in the southern part of the SXDS. In the zoomed-in circle, this overdense region, with $\delta_g \approx 3.8 \pm 0.7$, is outlined by the magenta half-circle-like shape. The blue and red points are the 46 spectroscopically confirmed LAEs in the two groups at $z \approx 5.68$ and $z \approx 5.75$, respectively (see also Fig. 3). The cyan rectangle represents SXDS_gPC, which is embedded in the large overdense region. It consists of 23 LAEs at $z \approx 5.68$ (blue points) in a near-square region of 15.5 × 14.5 arcmin^2, or ~37 × 35 cMpc^2. It has a high overdensity of $\delta_g \approx 5.6 \pm 1.2$.

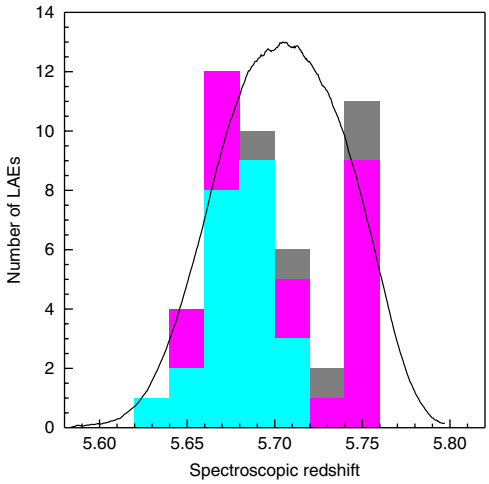

Fig. 3 | Redshift distribution of the LAEs. The dark grey, magenta and cyan histograms include all 46 LAEs. The magenta and cyan histograms together represent the 41 LAEs in the large overdense region, and the cyan histogram represents the 23 LAEs in SXDS_gPC only (see also Fig. 2). The black profile is the transmission curve of the NB816 filter used to select LAE candidates, scaled to a peak value of 13. The full-width at half-maximum of the filter is 120 Å, corresponding to $\Delta z \approx 0.10$ or a line-of-sight depth of ~46 cMpc. The histogram shows two peaks that represent two groups at $z \approx 5.68$ and $z \approx 5.75$, respectively.

redshift slice at $z \approx 5.75 \pm 0.01$. The two groups are shown in blue and red, respectively, in Fig. 2. SXDS_gPC is identified from the $z \approx 5.68$ group (see Methods for details). It consists of 23 LAEs in a near-square region of 15.5 × 14.5 arcmin^2 (~37 × 35 cMpc^2), denoted by the cyan rectangle in Fig. 2. SXDS_gPC has a high LAE density in a giant volume of ~35^3 cMpc^3, embedded in the even larger overdense region mentioned above. Its line-of-sight velocity dispersion is ~520 km s^{-1}, consistent with previously known large protoclusters at high redshift[12,13].

We measure galaxy overdensity $\delta_g \equiv n/\overline{n} - 1$, where n is the LAE number density in situ and $\overline{n}$ is the average density at $z \approx 5.7$. We calculate $\overline{n}$ on the basis of the LAEs found in our M2FS survey programme and the LAEs from two previous studies[20,21]. The two previous studies represent the two largest spectroscopic surveys of $z \approx 5.70$ LAEs made previously in fields other than the SXDS. They also used Suprime-Cam imaging data and the same NB816 filter for selection of LAE candidates. This substantially simplifies the calculation of δ_g because, in this case, a volume overdensity is equal to its corresponding surface overdensity. We count LAEs brighter than NB816 = 25.5 mag (the common magnitude limit in different samples) and correct for sample incompleteness. The resultant overdensities in the region surrounding and including SXDS_gPC are $\delta_g = 3.8 \pm 0.7$ and $\delta_g = 5.6 \pm 1.2$, respectively. This means that the LAE density in SXDS_gPC is 6.6 times the average density at $z \approx 5.7$. With the above LAEs used for the overdensity calculation, we estimate the significance of the overdensity, that is, statistically how significant the structure is in random fields. We find that $\delta_g = 5.6$ has a significance of ~5σ, indicating that the LAE overdensity in SXDS_gPC is highly significant.

We perform a test to demonstrate that SXDS_gPC is not a region with enhanced Lyα emission. We make use of the LAEs for the overdensity calculation above and estimate the fraction of LAEs brighter than 26.6 mag in the z' band for each sample. The z'-band photometry represents ultraviolet (UV) continuum flux because it does not cover the Lyα emission for the LAEs in this study. The adopted magnitude limit of 26.6 mag is roughly the 3σ detection limit for most z'-band images here. We find that this fraction in SXDS_gPC is very similar to those in other fields (see Methods for details). This clearly indicates that SXDS_gPC is not a region of increased Lyα emission relative to UV continuum emission. Instead, it is a highly overdense region of galaxies. The intrinsically high overdensity of SXDS_gPC exceeds the collapse threshold considerably in the classical theory

of spherical collapse. Cosmological simulations also suggest that an overdense region like SXDS_gPC will fall into a giant galaxy cluster[16].

Spectroscopically confirmed giant protoclusters like SXDS_gPC at $z>5$ have not been reported before. We estimate how rare they are by using cosmological simulations (see Methods for details). We update a previous work[16] that was based on the Millennium Run dark matter simulations[1]. We also incorporate a data-driven semi-analytic model[24] that represents one of the latest galaxy formation models. The simulation box size is 713 cMpc on each side. We use LAEs to trace dark matter, and search for protoclusters in a cubic window of 35^3 cMpc^3 at redshift near 5.7. The window size approximates the volume of SXDS_gPC. We find no giant protoclusters like SXDS_gPC with $\delta_g \approx 5.6$ in the entire simulation box. The highest overdensity found in the 35^3 cMpc^3 simulation windows is about 4.4. The probability of finding one system like SXDS_gPC in our survey area of 4 deg^2 is ~5%, implying that such systems are rare in the distant Universe.

We use two methods to estimate the present-day mass $M_{z=0}$, which is the total mass of baryonic matter and dark matter in SXDS_gPC (see Methods for details). We first use a classic formula[8], $M_{z=0} \approx (1+\delta_m)\overline{\rho}\,V$, where $\overline{\rho}$ is the current mean density of the Universe (3.88×10^{10} solar masses ($M_\odot$) per cMpc^3), V is the volume and δ_m is the mass overdensity. The value of δ_m is determined by $1+b\,\delta_m = C(1+\delta_g)$, where b is the bias parameter ($b=4.17$ from our simulation results above) and C is a small correction for redshift-space distortion. The resultant mass is $M_{z=0}=(3.4\pm0.6)\times10^{15}M_\odot$. This classic approach assumes that everything in the volume will collapse into a cluster, and the mass strongly depends on V. In contrast, this mass is not sensitive to the measured δ_g or b for a given V.

Our second method of mass measurement is to use the correlation between galaxy overdensity and present-day mass drawn from the simulation results[16]. This method does not require that everything in a protocluster volume has to fall into a cluster; thus, it is not sensitive to the assumed collapse volume. As we mentioned earlier, there are no systems like SXDS_gPC in our simulation results above, so we make use of the protocluster with the highest δ_g value, ~4.4, in the simulation. This protocluster has a present-day mass of $3.3\times10^{15}M_\odot$. On the basis of the classic formula above, for protoclusters with the same size at the same redshift, $M_{z=0}$ is proportional to $1+\delta_m$. Therefore, the mass ratio of SXDS_gPC to the $\delta_g \approx 4.4$ protocluster is the ratio of their $1+\delta_m$ values. From this calculation, we find that the present-day mass of SXDS_gPC is $\sim(3.6\pm0.9)\times10^{15}M_\odot$, consistent with the mass estimated from the classic formula. The two mass measurements demonstrate that SXDS_gPC is one of the most massive clusters or protoclusters currently known[25–27].

The cold dark matter model predicts that small structures merge hierarchically to form large structures, so the largest structures are expected to form in the latest cosmic times. It is thus remarkable that giant protoclusters such as SXDS_gPC already exist at $z\approx5.7$. Although SXDS_gPC is still far from virialized, its high overdensity suggests that this large overdense region must have been in place at an even earlier time. Such protoclusters may be ideal probes for understanding early structure formation. Furthermore, the discovery of SXDS_gPC has an intriguing implication for cosmic reionization, which ended at $z\approx6$. Some semi-analytic models and large-scale simulations have shown that the structure of reionization is mostly driven by the clustering of galaxies[28]. Under this picture, large-scale high-density regions were ionized first, where ionizing bubbles are thought to extend over tens of comoving megaparsecs[29,30]. The progenitor of SXDS_gPC is probably such a high-density region in the reionization era. Our results fit well into this scenario and may provide direct evidence for the existence of large-scale clustering required by the above reionization theory.

Methods

The M2FS survey programme. We are carrying out a spectroscopic survey of high-redshift galaxies in 4 deg^2 on the sky, using the fibre-fed multi-object spectrograph M2FS on the Magellan Clay Telescope. We aim to build a large and homogeneous sample of LAEs at $z\approx5.7$ and $z\approx6.5$ and Lyman-break galaxies at $5.5<z<6.8$. Our programme overview paper[23] presents the details of the programme, including the motivation, programme design, target selection, M2FS observations, LAE identification and scientific goals. The fields that we chose to observe are five well-studied fields, including the SXDS. These fields have deep optical imaging data in a series of broad and narrow bands, taken by Suprime-Cam on the 8.2 m Subaru Telescope. The combined SXDS images in five broad bands B, V, R, i' and z' have depths of 27.9 mag, 27.6 mag, 27.4 mag, 27.4 mag and 26.2 mag, respectively. The image in the narrow band NB816 reaches a depth of 26.1 mag.

Our main targets of the M2FS programme are LAEs at $z\approx5.7$ and $z\approx6.5$. We selected LAE candidates by using the narrow-band (or Lyα) technique. This technique has been extensively addressed in the literature, and selection criteria in different studies are similar. Here, we focus on the selection of $z\approx5.7$ LAEs in the SXDS. The selection was mainly based on the $i'-\mathrm{NB816}>1.0$ colour cut. We further required that candidates should not be detected ($<2\sigma$) in any bands bluer than the R band. We visually inspected each candidate and removed spurious detections caused by diffraction spikes from bright stars, cosmic rays, satellite trails and so on. We mainly considered relatively bright LAEs with NB816 ≤ 25.5 mag, but also included some slightly fainter candidates when possible.

Owing to a large field of view and high throughput, M2FS is very efficient for identifying relatively bright high-redshift galaxies. We use a pair of red-sensitive gratings with a resolving power of about 2,000. The wavelength coverage is roughly from 7,600 Å to 9,600 Å. The design of the M2FS pointings or plug plates is described in the overview paper[23]. For the SXDS, we use five M2FS pointings to cover most of its area (~1.2 deg^2), including one pointing covering SXDS_gPC and its surrounding region. Spectroscopic observations for our M2FS programme are ongoing. We have observed a total of ~2.5 deg^2 so far. The depth is not uniform, but many M2FS pointings have reached the required depth, including the pointing with SXDS_gPC.

LAEs in the SXDS_gPC region. The M2FS observations of the SXDS_gPC region were made during dark time in November 2015. From these observations, we confirmed 46 luminous LAEs at $z\approx5.70\pm0.05$ from their Lyα emission lines. The details of the 46 LAEs are provided in Supplementary Fig. 1 and Supplementary Table 1. The Lyα emission lines were identified and confirmed on the basis of both two-dimensional images and one-dimensional spectra. Our LAE selection criteria generally ensure that a detected line in the expected wavelength range around 8,160 Å is the Lyα emission line of an LAE at $z\approx5.7$. Four strong emission lines in star-forming galaxies that might contaminate our lines are [O II] λ3,727, Hβ, [O III] λ,5007 and Hα. The non-detection in very deep BVR images and the wavelength coverage of our spectral data rule out the possibility that the detected line is Hβ, [O III] λ,5007 or Hα. The most likely contaminants are [O II] emitters, as already pointed out in the literature. The [O II] line is a doublet, and our spectral resolution is high enough to resolve it. A tiny fraction of the candidates are indeed confirmed to be [O II] emitters at lower redshift. Finally, we can clearly see asymmetry in the emission lines of the confirmed LAEs. This is an indicator of the Lyα emission line at high redshift due to strong intergalactic-medium absorption blueward of the line.

Figure 2 illustrates the layout of the SXDS_gPC region. The whole SXDS field is scaled by 1/6 and is shown in dark grey in the figure. The black points represent spectroscopically confirmed LAEs brighter than NB816 = 25.5 mag. Their spatial distribution clearly indicates the existence of a large overdense region in the southern part of the SXDS. This is the overdense region that we identified from the photometric data. We use the magenta half-circle-like shape to denote this large overdense region. Most parts of its boundary are confined by the coverage of the imaging and spectroscopy available. Its upper (northeast) boundary is described as a straight line that is close to the two nearest LAEs in the overdense region. The projected size approximates to 22×17 arcmin^2, or 53×41 cMpc^2. The real size may be larger and beyond the coverage of the available data. This field was previously claimed[9] to have two compact overdense regions with eight secure LAEs at $z\approx5.7$. Their sizes were about several comoving megaparsecs, and their galaxy overdensities were estimated to be around 80. There has been no previous confirmation of giant protoclusters at $z\approx5.7$ like SXDS_gPC in this field.

Figure 3 shows the redshift distribution of the LAEs and the transmission curve of the NB816 filter. The starting point of the histogram is $z=5.62$ and the bin size is 0.02. The full-width at half-maximum of the filter is 120 Å, corresponding to $\Delta z\approx0.10$, or a line-of-sight depth of ~46 cMpc. Here, the line-of-sight depth derived from the redshift interval is a good approximation of the real line-of-sight distance. In the very early stage of the collapse, high-redshift protoclusters were just breaking away from the Hubble flow, so the redshift-space distortion is small.

Figure 3 clearly shows two distribution peaks at $z\approx5.68$ and $z\approx5.75$. The blue and red points in Fig. 2 represent the LAEs in the two groups, respectively. The majority of the LAEs are in the $z\approx5.68$ group. SXDS_gPC is identified from this group. In the literature, there is no clear definition for the boundary of a protocluster: previous studies have tended to draw a box (square or rectangle) or

a circle to include galaxies in a relatively isolated overdense region. Following this convention, we use a cyan rectangle in Fig. 2 to denote the SXDS_gPC boundary. Note that we have excluded five LAEs at right ascension of ~34.6 deg that are far away from other LAEs. We have also excluded three LAEs at declination of about −5.35 deg for the same reason. The remaining LAEs appear to be relatively isolated and have a near-square shape, on the basis of their spatial distribution. Therefore, we use the cyan rectangle to include these LAEs as the boundary of SXDS_gPC. The distance between each side of the rectangle and its nearest LAE is the median separation between the LAEs and their nearest neighbours. The projected size of SXDS_gPC is ~15.5×14.5 arcmin^2, or ~37×35 cMpc^2. This definition of the boundary is conservative. If the three LAEs at declination of about −5.35 deg were included, the projected size would increase by ~20% (but the galaxy overdensity would decrease by ~5%; see the next sections). The line-of-sight depth of SXDS_gPC is ~34 cMpc, so the volume is ~35^3 cMpc^3. We estimate the line-of-sight velocity dispersion from the spectroscopic redshifts by using a biweighted standard deviation. The measured value is ~520 km s^{-1}, consistent with those of other large protoclusters known at high redshift.

The overdensity in SXDS_gPC. In this section, LAEs are defined as LAEs brighter than 25.5 mag. As we mentioned earlier, we have observed a total of ~2.5 deg^2 for our M2FS programme. Although the depth is not uniform for the M2FS pointings observed so far, most pointings have reached the required depth. We find 86 $z \approx 5.70$ LAEs over a total of 5,160 arcmin^2 in these pointings (excluding the pointing with SXDS_gPC).

Two previous studies[20,21] used Suprime-Cam imaging data and the same NB816 filter for selection of $z \approx 5.70$ LAE candidates, meaning that a volume overdensity is equal to its corresponding surface overdensity because of the same line-of-sight depth. The Hu et al.[20] sample consists of seven individual fields, with a total coverage of 4,180 arcmin^2. More than 90% of their LAE candidates were spectroscopically observed down to a limit of NB816 ≈ 25.5 mag. They identified 73 LAEs. The Kashikawa et al.[21] sample contains one field, namely, the Subaru Deep Field (867 arcmin^2). More than 90% of their LAE candidates brighter than 25.5 mag have been spectroscopically observed (see their fig. 7). Most of the candidates were actually observed by Shimasaku et al.[19], who confirmed 28 LAEs brighter than 25.5 mag (ref. [19]).

To calculate the galaxy overdensity, we first correct for sample incompleteness. Sample incompleteness originates from four major sources: (1) object detection in imaging data, (2) galaxy candidate selection, (3) spectroscopic observations and (4) LAE identification in spectroscopic data. The first completeness is the probability of detecting an object in the NB816 images. The probability is roughly constant before the object brightness reaches a faint limit. The common limit for all samples used to calculate $\bar{n}$ is 25.5 mag (for example, fig. 3 in Hu et al.[20]). Therefore, we choose 25.5 mag as our limit, so that we do not need to apply corrections for this incompleteness.

The second source of sample incompleteness comes from target selection, that is, the colour cuts used to select targets. Different colour cuts select different fractions of real LAEs. The colour cut that we used is $i' - \mathrm{NB816} > 1.0$. We do not compute absolute completeness here. Instead, we apply small corrections so that each sample has the same colour cut of $i' - \mathrm{NB816} > 1.0$ (and therefore the same completeness). Hu et al.[20] used $I - \mathrm{NB816} > 0.8$. The filter I (centred at 794 nm) is slightly redder than i', and this cut is almost the same as our cut when the small difference is taken into account. Shimasaku et al.[19] also used the same criterion. Kashikawa et al.[21] used a more stringent cut of $i' - \mathrm{NB816} > 1.5$. We find that the fraction of LAEs with $i' - \mathrm{NB816} < 1.5$ is 16% in the combination of the Hu et al.[20] sample and our sample, so we apply a correction factor of 1/0.84.

The third source for sample incompleteness comes from spectroscopic observations, that is, the fraction of photometrically selected targets that were observed spectroscopically. This fraction is ~97% in our programme. We actually included nearly 100% LAEs in our M2FS observations, but we missed ~3% of the targets due to fibre problems (such as very low-efficiency fibres). We thus apply a correction factor of 1/0.97 to our sample. The Hu et al.[20] sample and Kashikawa et al.[21] sample have a completeness of ~90%, as we mentioned earlier. We apply a correction factor of 1/0.9 to the two samples.

The fourth major source for incompleteness comes from LAE identification in spectra. Because there are numerous OH skylines in the red part of the optical range, the fraction of LAEs that can be recovered from spectra is a function of wavelength for a given spectral resolution. However, the NB816 filter is located in an OH-dark window, so the effect of OH skylines is negligible. The completeness may also depend on instruments, observing modes, weather conditions and so on. As long as the spectroscopic data are deep enough, the completeness here is nearly 100%.

In summary, with the above incompleteness corrections, there are 73/0.9 LAEs over 4,180 arcmin^2 from the Hu et al.[20] sample, 28/0.9 LAEs over 867 arcmin^2 from the Kashikawa et al.[21] sample and 86/0.97 LAEs over 5,160 arcmin^2 from our sample. The average surface density is 0.0197 per arcmin^2. In the large overdense region, we have 35 LAEs brighter than NB816 = 25.5 mag in ~370 arcmin^2. The density is 0.095 per arcmin^2, about 4.8 times the average density. The resultant overdensity is $\delta_g = 3.8 \pm 0.7$, where the error is from Poisson statistics. Similarly, we have 22 LAEs brighter than NB816 = 25.5 mag in SXDS_gPC, and the overdensity is $\delta_g = 5.6 \pm 1.2$. Note that the line-of-sight depth of SXDS_gPC is 75% of that of the large overdense region.

Using the same LAEs used for the overdensity calculation above, we estimate the significance of the overdensity in SXDS_gPC, that is, statistically how significant the structure is compared with random fields. The Hu et al.[20] sample consists of seven individual fields, and the field size of 25×25 arcmin^2 is similar to the M2FS field of view. The Kashikawa et al.[21] sample is located in one field, and its size is slightly larger. We randomly place a 15.5×14.5 arcmin^2 cell (the size of SXDS_gPC) in all fields that we used above (including our own fields, but excluding the SXDS_gPC region), and count LAEs in the cell. A cell is not used if it crosses a field edge. This is performed 1,000 times. The resultant statistics show that the overdensity of SXDS_gPC is at ~5σ significance.

The cosmological simulation. We use a cosmological simulation to search for protoclusters at high redshift. This is an update of a previous work[16]. The simulation adopts the Planck cosmology and incorporates one of the latest galaxy formation models[24], which is a data-driven semi-analytic model. The simulation box size is 713 cMpc on each side. Galaxies selected on the basis of stellar mass or UV luminosity are often used as tracers of dark matter. However, our LAE sample was not selected by stellar mass or by UV luminosity. Instead, it is a Lyα-flux-limited sample (limited by the NB816-band flux). Therefore, we use LAEs as tracers of dark matter by linking Lyα emission to star formation rate (SFR).

We have followed the previous work[16] that used galaxies selected on the basis of SFR as tracers. We link Lyα emission to SFR by using an empirical relation in Jiang et al.[31] between Lyα-based SFR(Lyα) and UV-based SFR(UV). We also make the common assumption that SFR(UV) is the intrinsic SFR. For our sample, the SFR(Lyα) limit is ~$4\,M_\odot\,\mathrm{yr}^{-1}$, which corresponds to SFR(UV) ≈ $1\,M_\odot\,\mathrm{yr}^{-1}$ on the basis of the empirical relation mentioned above. We further test the limit of SFR(UV) by using the z'-band photometry. As seen from Supplementary Table 1, many LAEs are not significantly detected in the z' band. We visually identify the nine faintest LAEs that are barely detected in the z' band. We combine (average) them and perform a forced photometry on the expected position. The combined z'-band photometry is 28.7 mag. We assume a typical rest-frame UV slope of −2 and convert the z'-band photometry to SFR(UV), as done in Jiang et al.[31]. We find SFR(UV) = $0.92\,M_\odot\,\mathrm{yr}^{-1}$. This is well consistent with the above limit from SFR(Lyα). Therefore, we adopt $1\,M_\odot\,\mathrm{yr}^{-1}$ as the SFR limit of our sample.

We then use galaxies with SFR > $1\,M_\odot\,\mathrm{yr}^{-1}$ as tracers to search for protoclusters in the simulation. The LAE bias parameter at $z \approx 5.7$ derived from our simulation results is $b = 4.17$. This is well consistent with the $b = 4.11 \pm 0.17$ obtained from the recent results of the extensive photometric survey of $z \approx 5.70$ LAEs over ~14 deg^2 of the Subaru Hyper Suprime-Cam programme[32]. Note that the measured $M_{z=0}$ is not sensitive to the SFR limit that we adopt. The empirical relation between SFR(Lyα) and SFR(UV) has a scatter smaller than 50%. Even if we double the SFR limit to $2\,M_\odot\,\mathrm{yr}^{-1}$, the bias will increase slightly to $b = 4.64$ (an increase of ~10%; assuming it is an upper limit). On the basis of the classic formula used in the main text and in the section 'The mass of SXDS_gPC', $M_{z=0}$ is proportional to $1 + \delta_m$. An increase of b from 4.17 to 4.64 results in a ~5% decrease in the estimated $M_{z=0}$.

Finally, we perform a test to show that SXDS_gPC is not simply a region of increased Lyα emission relative to UV continuum emission. We make use of the LAEs that have been used to measure the galaxy overdensity in SXDS_gPC, and calculate the fraction of LAEs brighter than 26.62 mag in the z' band in each sample. The magnitude limit of 26.62 mag adopted here is the 3σ limit for the Kashikawa et al.[21] sample (see also Shimasaku et al.[19]). The 3σ limit of most z'-band images in our M2FS programme is similar to this limit. We exclude the fields that have much shallower z'-band images. The Hu et al.[20] sample does not have z'-band photometry available. In the Kashikawa et al.[21] sample, there are 9 out of 28 LAEs with z'-band detections (that is, brighter than 26.62 mag). After we correct for a sample incompleteness due to a slightly different colour selection criterion (see 'The overdensity in SXDS_gPC'), the fraction is $(9 + 28 \times 0.16)/28 \approx 48\%$. Strictly speaking, this fraction is the upper limit, assuming that all missed galaxies are brighter than the limit of 26.62 mag. In SXDS_gPC, the fraction of LAEs with $z' < 26.62$ mag is 57%. In the fields other than SXDS_gPC in our M2FS programme (excluding shallow regions mentioned above), the fraction is 53%. We can see that this fraction in SXDS_gPC is similar to, or even slightly higher than, those in other fields. This clearly indicates that SXDS_gPC is not a region with enhanced Lyα emission.

The mass of SXDS_gPC. We estimate the probability of finding one system like SXDS_gPC in our survey area of 4 deg^2 on the basis of the number of protoclusters found in the simulation. The expected number of protoclusters similar to SXDS_gPC in terms of volume and overdensity is equal to $\int n(\delta_g) P(\delta_g) \mathrm{d}\delta_g$, where $n(\delta_g)$ is the number of protoclusters with δ_g in the simulation, and $P(\delta_g)$ is the probability distribution of δ_g for SXDS_gPC. The resultant probability of finding one such system in 4 deg^2 is ~5%. Therefore, systems such as SXDS_gPC are rare at high redshift.

We use two methods to estimate the $M_{z=0}$ of SXDS_gPC. The details are described in the main text. The first one is the widely used formula $M_{z=0} \approx (1 + \delta_m)\bar{\rho}V$ (ref. [8]), where the value of δ_m is determined by $1 + b\,\delta_m = C(1 + \delta_g)$, where $C = 1 + f - f(1 + \delta_m)^{1/3}$ and $f = \Omega_m z^{4/7}$. The result is

$M_{z=0} = (3.4 \pm 0.6) \times 10^{15} M_{\odot}$, where the error includes the uncertainties of V, b and δ_g. This classic approach assumes that everything in the volume will collapse into a cluster, so the mass strongly depends on (is proportional to) V. As we can see from Fig. 2, we adopt a very conservative definition of the SXDS_gPC boundary that is nearly a lower limit. We assume a 10% relative uncertainty for V. In contrast, the mass is not sensitive to δ_g or b for a given V. On the basis of this classic formula, a 20% (~1σ uncertainty) increase of δ_g results in a ~10% increase in $M_{z=0}$, and a 10% increase (upper limit, assumed to be 1σ uncertainty) of b results in a 5% decrease in $M_{z=0}$.

In the second method, we estimate the present-day mass by comparing the simulation results on the basis of the relation between galaxy overdensity and present-day mass[16]. We use the $\delta_g = 4.4$ protocluster with $M_{z=0} = 3.3 \times 10^{15} M_{\odot}$ in the simulation. By scaling its mass, we obtain $M_{z=0} \approx (3.6 \pm 0.9) \times 10^{15} M_{\odot}$ for SXDS_gPC. The mass error is dominated by the uncertainty from the simulation. This uncertainty is the combination of the uncertainties that are reflected by the scatter in the relation between mass and galaxy overdensity. The mass error also includes the uncertainties of b and δ_g when we scale the mass of the $\delta_g = 4.4$ protocluster to the mass of SXDS_gPC. In addition, we assume a 10% relative uncertainty for V, as we did above. This method is based on a simulated correlation and does not require the whole protocluster region to entirely collapse. The derived mass is not sensitive to the assumed collapse volume V. For example, if we use a 45^3 cMpc^3 window size to search for protoclusters in the simulation, the volume size is increased by a factor of 2, but the mass is increased by only ~15%.

Data availability

The data that support the plots within this paper and other findings of this study are available from the corresponding author upon reasonable request.

Received: 16 April 2018; Accepted: 4 September 2018;
Published online: 15 October 2018

References

1. Springel, V. et al. Simulations of the formation, evolution and clustering of galaxies and quasars. *Nature* **435**, 629–636 (2005).
2. Kravtsov, A. V. & Borgani, S. Formation of galaxy clusters. *Ann. Rev. Astron. Astrophys.* **50**, 353–409 (2012).
3. Papovich, C. et al. A Spitzer-selected galaxy cluster at $z = 1.62$. *Astrophys. J.* **716**, 1503–1513 (2010).
4. Gobat, R. et al. A mature cluster with X-ray emission at $z = 2.07$. *Astron. Astrophys.* **526**, 133–145 (2011).
5. Stanford, S. A. et al. Discovery of a massive, infrared-selected galaxy cluster at $z = 1.75$. *Astrophys. J.* **753**, 164–171 (2012).
6. Andreon, S. et al. JKCS041: a Coma cluster progenitor at $z = 1.803$. *Astron. Astrophys.* **565**, 120–134 (2014).
7. Wang, T. et al. Discovery of a galaxy cluster with a violently starbursting core at $z = 2.506$. *Astrophys. J.* **828**, 56–70 (2016).
8. Steidel, C. C. et al. A large structure of galaxies at redshift z ~ 3 and its cosmological implications. *Astrophys. J.* **492**, 428–438 (1998).
9. Ouchi, M. et al. The discovery of primeval large-scale structures with forming clusters at redshift 6. *Astrophys. J. Lett.* **620**, 1–4 (2005).
10. Venemans, B. P. et al. Protoclusters associated with $z > 2$ radio galaxies—I. Characteristics of high redshift protoclusters. *Astron. Astrophys.* **461**, 823–845 (2007).
11. Capak, P. L. et al. A massive protocluster of galaxies at a redshift of z≈5.3. *Nature* **470**, 233–235 (2011).
12. Toshikawa, J. et al. Discovery of a protocluster at z ~ 6. *Astrophys. J.* **750**, 137–148 (2012).
13. Dey, A. et al. Spectroscopic confirmation of a protocluster at z ≈ 3.786. *Astrophys. J.* **823**, 11–28 (2016).
14. Franck, J. R. & McGaugh, S. S. The Candidate Cluster and Protocluster Catalog (CCPC)-II. Spectroscopically identified structures spanning $2 < z < 6.6$. *Astrophys. J.* **833**, 15–33 (2016).
15. Overzier, R. A. et al. ΛCDM predictions for galaxy protoclusters—I. The relation between galaxies, protoclusters and quasars at $z \sim 6$. *Mon. Not. R. Astron. Soc.* **394**, 577–594 (2009).
16. Chiang, Y.-S., Overzier, R. & Gebhardt, K. Ancient light from young cosmic cities: physical and observational signatures of galaxy protoclusters. *Astrophys. J.* **779**, 127–142 (2013).
17. Furusawa, H. et al. The Subaru XMM-NEWTON Deep Survey (SXDS)—II. Optical imaging and photometric catalogs. *Astrophys. J. Suppl.* **176**, 1–18 (2008).
18. Rhoads, J. E. et al. A luminous Lyα-emitting galaxy at redshift $z = 6.535$: discovery and spectroscopic confirmation. *Astrophys. J.* **611**, 59–67 (2004).
19. Shimasaku, K. et al. Lyà emitters at z=5.7 in the Subaru Deep Field. *Publ. Astron. Soc. Jpn* **58**, 313–334 (2006).
20. Hu, E. M. et al. An atlas of $z = 5.7$ and $z = 6.5$ Lyα emitters. *Astrophys. J.* **725**, 394–423 (2010).
21. Kashikawa, N. et al. Completing the census of Lyα emitters at the reionization epoch. *Astrophys. J.* **734**, 119–137 (2011).
22. Mateo, M. et al. M2FS: the Michigan/Magellan Fiber System. *Proc. SPIE* **8446**, 84464Y (2012).
23. Jiang, L. et al. A Magellan M2FS spectroscopic survey of galaxies at $5.5 < z < 6.8$: program overview and a sample of the brightest Lyα emitters. *Astrophys. J.* **846**, 134–148 (2017).
24. Henriques, B. M. B. et al. Galaxy formation in the Planck cosmology—I. Matching the observed evolution of star formation rates, colours and stellar masses. *Mon. Not. R. Astron. Soc.* **451**, 2663–2680 (2015).
25. Menanteau, F. et al. The Atacama Cosmology Telescope: ACT-CL J0102–4915 'El Gordo', a massive merging cluster at redshift 0.87. *Astrophys. J.* **748**, 7–24 (2012).
26. Casey, C. M. et al. A massive, distant proto-cluster at $z = 2.47$ caught in a phase of rapid formation. *Astrophys. J. Lett.* **808**, 33–40 (2015).
27. Cai, Z. et al. MApping the Most Massive Overdensities Through Hydrogen (MAMMOTH)-II. Discovery of the extremely massive overdensity BOSS1441 at z=2.32. *Astrophys. J.* **839**, 131–141 (2017).
28. Chiang, Y.-K. et al. Galaxy protoclusters as drivers of cosmic star-formation history in the first 2 Gyr. *Astrophys. J. Lett.* **844**, 23–29 (2017).
29. Wyithe, J. S. B. & Loeb, A. A characteristic size of ~10 Mpc for the ionized bubbles at the end of cosmic reionization. *Nature* **432**, 194–196 (2004).
30. Iliev, I. T. et al. Simulating cosmic reionization at large scales—I. The geometry of reionization. *Mon. Not. R. Astron. Soc.* **369**, 1625–1638 (2006).
31. Jiang, L. et al. Physical properties of spectroscopically confirmed galaxies at z≥6—I. Basic characteristics of the rest-frame UV continuum and Lyα emission. *Astrophys. J.* **772**, 99–118 (2013).
32. Ouchi, M. et al. Systematic Identification of LAEs for Visible Exploration and Reionization Research Using Subaru HSC (SILVERRUSH)—I. Program strategy and clustering properties of ~2,000 Lyα emitters at z=6–7 over the 0.3–0.5 Gpc^2 survey area. *Publ. Astron. Soc. Jpn* **70**, S13 (2018).

Acknowledgements

We acknowledge support from the National Key R&D Program of China (2016YFA0400703 and 2016YFA0400702) and from the National Science Foundation of China (grant 11533001). G.A.B. is supported by CONICYT/FONDECYT, Programa de Iniciacion, Folio 11150220. E.W.O. acknowledges support from the NSF from grant AST1313006. We thank R. de Grijs and M. Ouchi for discussions. This paper includes data gathered with the 6.5 m Magellan Telescopes located at Las Campanas Observatory, Chile. Australian access to the Magellan Telescopes was supported through the National Collaborative Research Infrastructure Strategy of the Australian Federal Government.

Author contributions

L.J. is the Principal Investigator of the project, analysed the data and prepared the manuscript. J.W. reduced the M2FS images. F.B., Y.S., Z.-Y.Z., J.I.B., J.D.C., M.M. and E.W.O. helped with the M2FS observations. Y.-K.C. carried out the simulations. L.C.H., X.F., R.W. and X.-B.W. prepared the manuscript. G.A.B. and G.A.O. helped with the M2FS data reduction. All authors helped with the scientific interpretations and commented on the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41550-018-0587-9.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to L.J.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.