Galaxies at High Redshift and Their Evolution Over Cosmic Time
Proceedings IAU Symposium No. 319, 2015
S. Kaviraj, ed.

doi:10.1017/S1743921315010911

Galaxy morphologies in the era of big-data surveys

M. Huertas-Company

Observatoire de Paris - Université Paris Diderot

Abstract. Galaxy morphology is a first-order descriptor of a galaxy and a useful proxy to identify physical processes. The 100 years old Hubble fork describes the structural diversity of galaxies in the local universe. Unveiling the origins of this galaxy zoology is a key challenge in galaxy evolution. In this review talk, I first summarized some key advances in our understanding of the morphological evolution of galaxies from $z \sim 0$ to $z \sim 3$, thank you in particular to the SDSS and HST legacies. In the second part, I focused on the classification techniques. With the emergence in the last years of large surveys the samples of study have increased by several orders of magnitude going from a few tens to several millions of objects. This trend will clearly continue in the next decade with coming surveys/missions such as EUCLID and WFIRST. While galaxy classification is still a required step in any survey, visual inspection of galaxies is becoming prohibitively time-consuming. Under these circumstances, the techniques used to estimate galaxy morphologies need to be updated.

Keywords. galaxy evolution, morphologies, machine learning etc.

1. Introduction

Whenever the human brain is faced to a complex problem, the very first approach is to group objects into groups of similar appearance (morphology). The basic assumption is that objects that look similar have somehow experienced the same evolution. This approach has some obvious caveats. For example, morphological convergence is a well known effect in biology that makes species with very different histories end up with a similar appearance. However, at first order, it represents a good way to approach a new problem. In astronomy, this effort was carried out almost 100 years ago by the american astronomer E. Hubble and the resulting classification scheme is known as the Hubble sequence (Hubble 1926, Figure 1). This very first optical classification divided galaxies into two main types based on the presence or not of a disk and revealed that the underlying population of galaxies in the local Universe is in fact bimodal. Despite many revisits, the Hubble Fork is still alive. At the point that, understanding the physical processes that lead to such a bimodality - i.e. how bulges and disks form and evolve is one of the major challenges in the field of galaxy evolution and the main goal of deep field surveys. High quality multi-wavelength data at different redshifts have enabled to establish the link between morphology and the physical properties of galaxies at different cosmic epochs. Despite of these significant progresses, the classification of galaxies at different cosmic epochs is still an important first-order descriptor and a good channel to trigger physics. In that respect, the increase of the amount of data has forced the community to find alternatives to the classical visual approach. The next-generation of big-data surveys such as EUCLID, LSST or WFIRST represent a new challenge in that respect for which no valid solution has been provided yet.

My talk at the General Assembly was divided in two main parts which I translate into this summary. First, I review the major advances in our understanding of how (massive

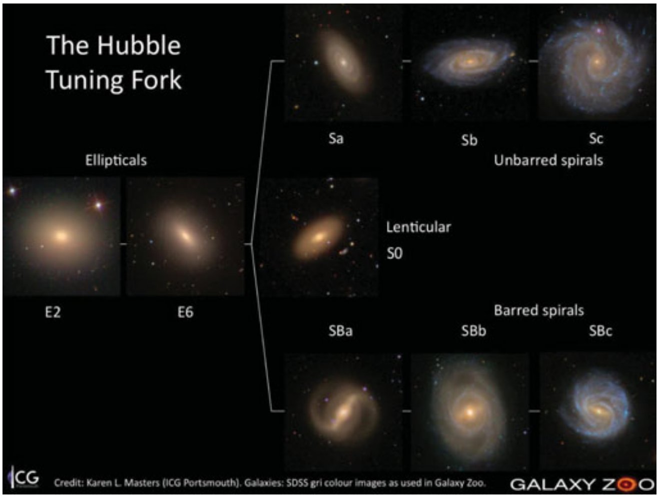

Figure 1. Hubble Sequence. [Credit: K. Masters - GALAXYZOO]

- $logM_* > 10$) galaxies change their morphology and secondly I focus on the evolution of the techniques used to estimate those morphologies in the era of large surveys.

2. The evolution of the Hubble sequence from $z \sim 3$

2.1. *The local Hubble sequence (the SDSS legacy)*

The Sloan Digital Sky Survey and its multiple follow-ups have enabled to acquire a reasonable knowledge of the Hubble sequence properties in our present universe. They have confirmed in fact that the bimodality first observed in the shapes of galaxies is indeed translated into diverse physical properties. This justifies the use of galaxy morphology as a first-order proxy for physical mechanisms. The abundance of each morphological type, as measured by the luminosity and stellar mass functions is well constrained. Bulge-dominated galaxies dominate the high-mass end of the SMF while later types tend to be more abundant at lower masses. The properties of the stellar populations are also well distinguished. It is now well established that early-type galaxies are dominated by old stellar populations (e.g. Brinchmann *et al.* 2004). The morphology-density relation established that the morphological mixing of galaxy populations is also modulated by environment. Different Hubble types also follow different scaling relations (e.g. mass-size relation - Bernardi *et al.* 2014), witnessing a different dynamical status as also probed by integral field spectroscopic surveys. Unveiling the emergence of this structural, morphological, kinematical and chemical bimodality is one of the main challenges in galaxy evolution.

2.2. *The Hubble sequence at high redshift (the HST legacy)*

A major step was given in the late 90's when the Hubble Space Telescope opened the window to the distant universe with the Hubble Deep Field. The most striking result of this first inspection was the significant increase of the abundance of *irregular* systems (e.g. Abraham *et al.* 1996; Conselice *et al.* 2000) , which are almost inexistent in the local universe (above 10^{10} solar masses). Since then, multiple deep imaging surveys (GOODS, DEEP2, COSMOS, CANDELS etc) have mapped and quantified the morphological content of the universe from $z \sim 3$, confirming that at $z > 1$, the vast majority of galaxies present an irregular distribution of their light profiles (Figure 2). These star-forming

https://doi.org/10.1017/S1743921315010911 Published online by Cambridge University Press

disturbed morphologies coexist with compact massive spheroids (e.g. Trujillo *et al.* 2006) that are present at very early epochs and which formation and evolution has been widely debated in the last years (e.g. Newman *et al.* 2012). The symmetric and regular disks that populate our surrounding universe today, were hence a minority ~ 10 Gyrs ago. Again, as it happened in the local universe, this first inspection of the morphologies triggered a variety of works focused on understanding this diversity and the subsequent evolution. The obvious interpretation was that galaxies appeared disturbed because they were gravitationally interacting with other galaxies (merger event). Mergers are indeed a channel to build bulges (it was known since the 70's- Toomre & Toomre 1972) and even late type spirals if a disk rebuilding event from the surrounding gas follows the merger (e.g. Hammer *et al.* 2005). However, multiple follow-up observations have changed this initial picture (this highlights the limits of galaxy morphology to establish physics). The inferred major merger fractions do not seem to be enough to explain the abundance of bulges in the local universe despite of large uncertainties (e.g. Lotz *et al.* 2011). Also, star formation in galaxies has been observed to be surprisingly regulated, in the sense that the specific star formation rate in galaxies is remarkably constant at a given epoch (i.e. main sequence of star formation - Brinchmann *et al.* 2004). This is interpreted as an indirect evidence that fuel in the form of cold gas is somehow continuously being fed into the galaxies to sustain star formation (minor mergers and/or gas accretion, e.g. Dekel *et al.* 2009), as opposed to violent stochastic events expected from major mergers. Kinematical studies of normal main sequence galaxies at $z > 1$ have confirmed in fact that, despite their small size and disturbed light profiles, their gas and stars are rotating as present day spirals for most of them (e.g. Wisnioski *et al.* 2015). The fundamental difference between local and distant galaxies seems to reside in the star formation efficiency which was significantly higher in the early universe (Daddi *et al.* 2007) and cannot be only explained by the higher gas fraction, and the high turbulence of the inter-galactic medium (e.g. Genzel *et al.* 2008). The picture is therefore emerging that during most of its life, a typical massive galaxy seems to live a rather quiet life. How does the measured galaxy bimodality emerges then? Two major events, eventually related, can break this apparent equilibrium. An episode of high star formation activity (e.g starburst) can be triggered (e.g. Elbaz *et al.* 2002). Or, suddenly something might happen that prevents the galaxy to continue forming new stars (e.g. Peng *et al.* 2010). This process is known as quenching, and provokes that the galaxy looses its population of blue newly-born stars and becomes dominated by red old ones. Quenching, seems to be the fundamental mechanism that helps explaining many of the properties of our surrounding universe, and in particular the galaxy morphological evolution. There is indeed a measured correlation between the galaxy star-formation rate and its morphology and structure from $z \sim 3$ (e.g Wuyts *et al.* 2011; Huertas-Company *et al.* 2015). Whether these correlations are explained by a consequence or causal connection is still being debated (e.g. Carollo *et al.* 2014; Barro *et al.* 2015)

Why a galaxy would quench? Given that star-formation is related to cold gas, a galaxy can only quench, either because 1) it runs out of cold gas or 2) because something prevents it to cool efficiently. Several mechanisms can help removing gas from the galaxy. Sudden energy/momentum release from star formation and/or AGN (feedback) results in the ejection of all gas from galaxies (e.g. Granato *et al.* 2004). Also encounters with other galaxies can efficiently remove gas (tidal, ram pressure). On the other side, the gas reservoir can also be slowly depleted through stellar evolution (secular) provided that no more cold gas enters the system (e.g. Gavazzi *et al.* 2015). This can be achieved for example if the circumgalactic gas is shock-heated to high temperatures as the mass of the host dark matter halo exceeds a critical threshold (of order of $10^{12} M_{\odot}$), and therefore

https://doi.org/10.1017/S1743921315010911 Published online by Cambridge University Press

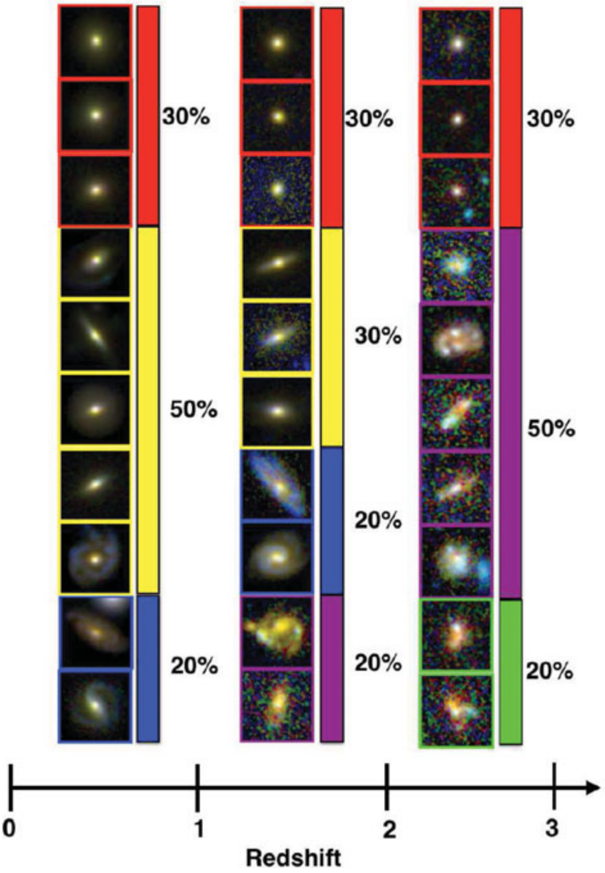

Figure 2. Fraction of different optical rest-frame morphologies from $z \sim 3$ for typical $10^{11} M_*/M_\odot$ progenitors. Objects are selected suing abundance matching to limit the progenitor bias effect. Adapted from Huertas-Company *et al.* (2015)

it stops to cool and flow into the galaxy (e.g. Birnboim & Dekel 2003). Powerful AGN jets may also heat gas to high temperature thus preventing further accretion of cold gas (radio mode AGN feedback). Finally, the growth of a central mass concentration (bulge) may suppress the disc instability and prevent the formation of star-forming clumps, the so-called morphological (or gravitational) quenching (Martig *et al.* 2009). So there are a variety of different mechanisms actually considered for the quenching process. Which one is dominantly driving galaxy evolution (if there is) or under which circumstances one or another process is triggered is still a mystery and a fundamental question to be addressed in the following years.

2.3. *JWST and big-data surveys*

The field of galaxy morphologies will clearly evolve in two main fronts in the next decade. On the one hand, facilities like JWST will enable to probe the universe at $z > 3$, currently not accessible by HST on a statistical basis. Also the low mass end of the galaxy distribution at high redshift will be unveiled. As it happened in the 90's with Hubble, we expect the discovery of new objects which cannot be observed with current facilities. In that respect, morphology will be again the fist order quantity to be measured to trigger some new (?) physics.

https://doi.org/10.1017/S1743921315010911 Published online by Cambridge University Press

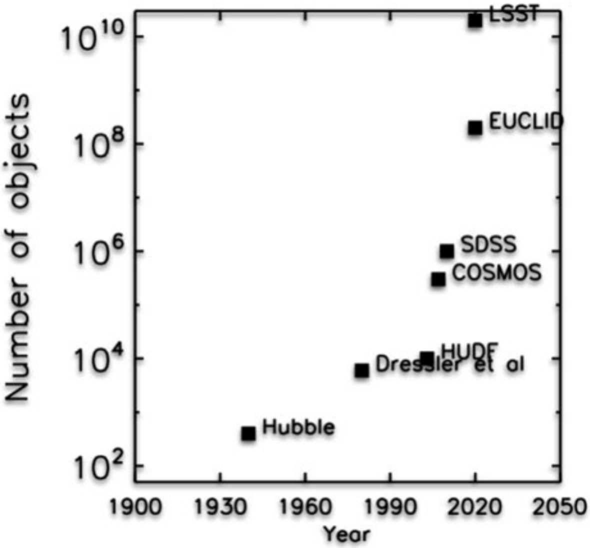

Figure 3. Number of observed objects in some key surveys asa function of time. The figure is not complete but reflects the clear increasing trend.

The other interesting and new front is the so-called big-data revolution in astronomy. Surveys of billions of objects are planed for the next 2-5 years. Extracting and interpreting the information contained in these big-data surveys represent a new challenge for which no valid solution exists today. In the second part of this short review, I summarize the efforts done by the community to cope with the data volume increase.

3. Estimating galaxy morphologies in large surveys

3.1. *Citizen science*

One beautiful solution has been to dramatically increase the number of involved people through what has been called *citizen science*. The Galaxy Zoo (GZOO) project (Lintott *et al.* 2011) involves more than 600.000 people all around the world to morphologically classify the full SDSS sample and is now been extended to other higher redshift samples. A similar approach has been followed in high redshift surveys such as CANDELS, but only with professional astronomers (Kartaltepe *et al.* 2015). The obvious advantage of such an approach is that detailed morphologies can be estimated without inventing new algorithms while still taking advantage of the human brain inherent capacities to detect complex features in images. There are some obvious problems though such as the organization, the necessity to train non-expert people (in the case of the GZOO) but also the non-reproducibility. But above all, with the next generation of surveys we are probably reaching as well the limit of applicability of these approaches. First estimations reveal that we would need close to a hundred years to classify all data from the ESA mission EUCLID with a Galaxy Zoo like approach unless the amount of involved people is significantly increased.

3.2. *CAS-based methods*

A question naturally arises, can we train computers to do the job or at least to help? There have been some efforts led by different groups towards that direction consisting on using existing visual morphologies on a smaller dataset to train automated machine learning

https://doi.org/10.1017/S1743921315010911 Published online by Cambridge University Press

algorithms (e.g. Huertas-Company *et al.* 2008; Ball *et al.* 2004). The basic idea behind these approaches is to find a set of parameters which somehow correlate with the visual morphology of a galaxy and compute some thresholds in the parameter space that allow to identify the different morphological types (e.g Abraham *et al.* 1996; Conselice *et al.* 2000). In astronomy, these parameters traditionally include concentrations, asymmetries, clumpiness (or smoothness), gini coefficient, moments of light etc (e.g. Abraham *et al.* 2003; Conselice *et al.* 2000; Lotz *et al.* 2004). In the last years, we proposed a generalization of this approach with the development of galSVM (Huertas-Company *et al.* 2008, 2011), which enables an n-dimension classification with optimal non-linear boundaries in the parameter space as well as a quantification of errors following a probabilistic approach (see also Scarlata *et al.* 2007). These *CAS (Concentration-Asymmetry-Smoothness) based* methods have been proved to be relatively useful but are also faced to several limitations. The values of the parameters strongly depend on the data quality and redshift and they only provide with rough morphological classifications in 2 or 3 classes. The most important problem with such techniques is perhaps that the fraction of miss-classifications is still high at high redshift specially ($\sim$20 − 30%, Huertas-Company *et al.* 2014).

3.3. *Deep-Learning: a promising tool for big-data surveys*

The problem might reside in the parameters that are traditionally used. Concentrations, asymmetries etc are useful because they reduce the complexity of the problem by describing a galaxy with just a few parameters, but also imply that a tremendous amount of information contained in the pixels themselves is actually lost. As a consequence they might not be necessarily well adapted to what the human brain actually does which is looking at the full distribution of light. Interestingly, the big-data revolution in many different fields together with the advent of powerful computing resources such as GPUs, has enabled the development and application of new learning techniques that use all the pixels as parameter space. They are by this fact more suited to mimic the human perception. Deep-learning (DL) is indeed a non-linear learning process that automatically learns and extracts the most relevant features for the problem it is being trained to solve. The key point of DL is that it does not assume any a priori knowledge of the parameters that need to be plugged into the network. Instead, it learns them in a non-linear way and selects the optimal features that best correlate with the quantity to be predicted. Though deep learning architectures have existed since the early 80s, they involve complex technological problems that only allowed their use in real problems in the last few years. The main limitation is that the feature learning process needs to be performed in large enough datasets to prevent over-fitting because of the large number of free parameters that the model contains. The availability of extra large datasets opens the window to the application of such methods. DL is hence not only a statistical tool, but it also has the ability to reduce the systematic uncertainties taking advantage of large statistics.

Deep learning was first applied to galaxy morphology earlier this year in the framework of an online competition set up by the galaxy zoo team. The driving idea was to ask the machine learning community to find an algorithm that best reproduces the Galaxy Zoo classifications using a training set of $\sim$ 20.000 galaxies. The winner of the competition applied a convolution neural network to reach an root mean square error of $\sim$ 7% (Dieleman *et al.* 2015) in all the 37 features measured in the GZOO. In a more recent work, we tested this approach on high redshift galaxies ($1 < z < 3$) from the CANDELS survey (see Figure 4, Huertas-Company *et al.* 2015). Using a visual training set performed by professional astronomers from the CANDELS collaboration (Kartaltepe *et al.* 2015), we obtained a $\sim$ 95% agreement between visual and deep-learning based classifications in 5 broad morphological classes (spheroids, disks, irregulars, point sources and

https://doi.org/10.1017/S1743921315010911 Published online by Cambridge University Press

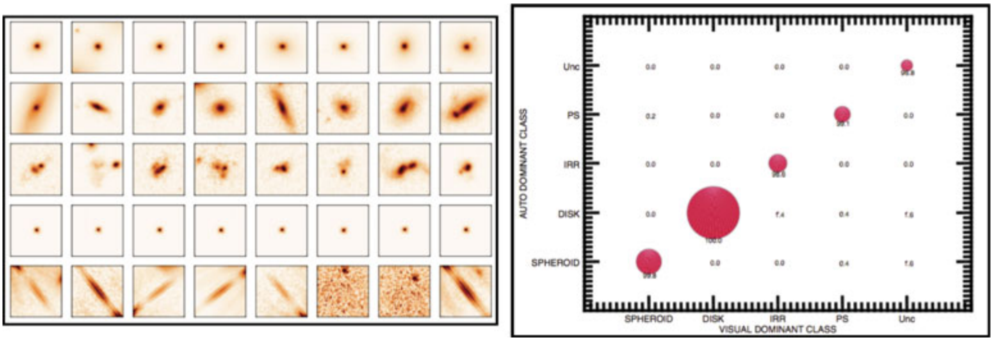

Figure 4. Example of deep learning for bulge detection applied to CANDELS. Left: Galaxies are classified in 5 broad classes. From top to bottom: spheroids, disk +spheroids, irregulars, point-sources and bad detections. Right: Accuracy of the classification compared to a human based inspection. We reach an agreement $> 95\%$. [Adapted from Huertas-Company *et al.* (2015)]

unclassifiable). This represents a major improvement compared to previous CAS based methods combined with a machine learning layer and therefore a very promising tool for the future. It illustrates nicely how the combination of citizen science based approaches with powerful intelligent algorithms might be the way to go.

4. Summary and conclusions

Morphology continues to be a solid tracer of physics. From $z \sim 0$ to $z \sim 3$, the morphologies of galaxies do correlate with their physical properties (star-formation, stellar-populations etc). The exact physical mechanisms leading to these correlations are still debated. Thank you to large surveys with Hubble Space Telescope imaging at different wavelengths, we have now a reasonable view of how the morphologies of massive galaxies changed with time. The local Hubble sequence seems to be in place at $z \sim 1$, while at higher redshifts it is completely dominated by disturbed morphologies. A fraction of massive compact spheroids already exist at very early epochs. Kinematic studies of typical star-forming galaxies have shown though, that despite their appearance, they are mostly rotating at these redshifts and therefore major mergers do not seem to be the main channel to explain the morphological evolution of galaxies. Other mechanisms such as bulge growth through disk instabilities followed by morphological quenching need to be included in the puzzle.

Estimating galaxy morphologies in large surveys is still a challenge for which no standard solution has been provided yet. The citizen science approach provides a good channel to benefit from the efficiency of the human brain to detect complex features. There are however some obvious problems to its generalization such as the amount of time required and the lack of reproductivity. On the other hand, machine based techniques do not usually offer a good enough trade-off between accuracy and details. Deep-learning techniques allow to combine both approaches an appears to be a promising option for future surveys that needs to be explored into more details.

References

Abraham, R. G., van den Bergh, S., & Nair, P. 2003, *ApJ*, 588, 218
Abraham, R. G., van den Bergh, S., Glazebrook, K., *et al.* 1996, *ApJS*, 107, 1
Ball, N. M., Loveday, J., Fukugita, M., *et al.* 2004, *MNRAS*, 348, 1038
Barro, G., Faber, S. M., Koo, D. C., *et al.* 2015, arXiv:1509.00469

https://doi.org/10.1017/S1743921315010911 Published online by Cambridge University Press

Bernardi, M., Meert, A., Vikram, V., *et al.* 2014, *MNRAS*, 443, 874
Birnboim, Y. & Dekel, A. 2003, *MNRAS*, 345, 349
Brinchmann, J., Charlot, S., White, S. D. M., *et al.* 2004, *MNRAS*, 351, 1151
Carollo, C. M., Cibinel, A., Lilly, S. J., *et al.* 2014, arXiv:1402.1172
Conselice, C. J., Bershady, M. A., & Jangren, A. 2000, *ApJ*, 529, 886
Daddi, E., Dickinson, M., Morrison, G., *et al.* 2007, *ApJ*, 670, 156
Dekel, A., Birnboim, Y., Engel, G., *et al.* 2009, *Nature*, 457, 451
Dieleman, S., Willett, K. W., & Dambre, J. 2015, *MNRAS*, 450, 1441
Elbaz, D., Cesarsky, C. J., Chanial, P., *et al.* 2002, *A&A*, 384, 848
Gavazzi, G., Consolandi, G., Dotti, M., *et al.* 2015, *A&A*, 580, A116
Genzel, R., Burkert, A., Bouché, N., *et al.* 2008, *ApJ*, 687, 59
Granato, G. L., De Zotti, G., Silva, L., Bressan, A., & Danese, L. 2004, *ApJ*, 600, 580
Hammer, F., Flores, H., Elbaz, D., *et al.* 2005, *A&A*, 430, 115
Hubble, E. P. 1926, *ApJ*, 64, 321
Huertas-Company, M., Pérez-González, P. G., Mei, S., *et al.* 2015, *ApJ*, 809, 95
Huertas-Company, M., Gravet, R., Cabrera-Vives, G., *et al.* 2015, *ApJS*, 221, 8
Huertas-Company, M., Kaviraj, S., Mei, S., *et al.* 2014, arXiv:1406.1175
Huertas-Company, M., Aguerri, J. A. L., Bernardi, M., Mei, S., & Sánchez Almeida, J. 2011, *A&A*, 525, A157
Huertas-Company, M., Rouan, D., Tasca, L., Soucail, G., & Le Fèvre, O. 2008, *A&A*, 478, 971
Kartaltepe, J. S., Mozena, M., Kocevski, D., *et al.* 2015, *ApJS*, 221, 11
Lintott, C., Schawinski, K., Bamford, S., *et al.* 2011, *MNRAS*, 410, 166
Lotz, J. M., Jonsson, P., Cox, T. J., *et al.* 2011, *ApJ*, 742, 103
Lotz, J. M., Primack, J., & Madau, P. 2004, *AJ*, 128, 163
Martig, M., Bournaud, F., Teyssier, R., & Dekel, A. 2009, *ApJ*, 707, 250
Newman, A. B., Ellis, R. S., Bundy, K., & Treu, T. 2012, *ApJ*, 746, 162
Peng, Y.-j., Lilly, S. J., Kovač, K., *et al.* 2010, *ApJ*, 721, 193
Scarlata, C., Carollo, C. M., Lilly, S., *et al.* 2007, *ApJS*, 172, 406
Toomre, A. & Toomre, J. 1972, *ApJ*, 178, 623
Trujillo, I., Feulner, G., Goranova, Y., *et al.* 2006, *MNRAS*, 373, L36
Wisnioski, E., Förster Schreiber, N. M., Wuyts, S., *et al.* 2015, *ApJ*, 799, 209
Wuyts, S., Förster Schreiber, N. M., van der Wel, A., *et al.* 2011, *ApJ*, 742, 96

https://doi.org/10.1017/S1743921315010911 Published online by Cambridge University Press