Maximum parsimony analysis of the effect of the environment on the evolution of galaxies

Didier Fraix-Burnet¹, Mauro D’Onofrio², and Paola Marziani³

¹ Univ. Grenoble Alpes, CNRS, IPAG, Grenoble, France

e-mail: didier.fraix-burnet@univ-grenoble-alpes.fr

² Dipartimento di Fisica & Astronomia, Università di Padova, Italy

e-mail: mauro.donofrio@unipd.it

³ INAF, Osservatorio Astronomico di Padova, Italy

e-mail: paola.marziani@inaf.it

Received 3 April 2019 / Accepted 25 July 2019

ABSTRACT

Context. Galaxy evolution and the effect of the environment are most often studied using scaling relations or regression analyses around a given property. However, these approaches do not take into account the complexity of the physics of the galaxies and their diversity.

Aims. We here investigate the effect of the cluster environment on the evolution of galaxies through multivariate, unsupervised classification and phylogenetic analyses applied to two relatively large samples from the Wide-field Nearby Galaxy-cluster Survey (WINGS), one of cluster members and one of field galaxies (2624 and 1476 objects, respectively).

Methods. These samples are the largest ones ever analysed with a phylogenetic approach in astrophysics. To be able to use the maximum parsimony (cladistics) method, we first performed a pre-clustering in 300 clusters with a hierarchical clustering technique, before applying it to these pre-clusters. All these computations used seven parameters: $B-V$, $\log(R_e)$, n_V , $\langle\mu\rangle_e$, H_β , D_{4000} , and $\log(M^*)$.

Results. We have obtained a tree for the combined samples and do not find different evolutionary paths for cluster and field galaxies. However, the cluster galaxies seem to have accelerated evolution in the sense that they are statistically more diversified from a primitive common ancestor. The separate analyses show a hint of a slightly more regular evolution of the variables for the cluster galaxies, which may indicate they are more homogeneous compared to field galaxies in the sense that the groups of the latter appear to have more specific properties. On the tree for the cluster galaxies, there is a separate branch that gathers rejuvenated or stripped-off groups of galaxies. This branch is clearly visible on the colour-magnitude diagram, going back from the red sequence towards the blue one. On this diagram, the distribution and the evolutionary paths of galaxies are strikingly different for the two samples. Globally, we do not find any dominant variable able to explain either the groups or the tree structures. Rather, co-evolution appears everywhere, and could depend itself on environment or mass.

Conclusions. This study is another demonstration that unsupervised machine learning is able to go beyond simple scaling relations by taking into account several properties together. The phylogenetic approach is invaluable in tracing the evolutionary scenarios and projecting them onto any bivariate diagram without any a priori modelling. Our WINGS galaxies are all at low redshift, and we now need to go to higher redshifts to find more primitive galaxies and complete the map of the evolutionary paths of present day galaxies.

Key words. galaxies: evolution – galaxies: clusters: general – galaxies: statistics – methods: data analysis – methods: statistical

1. Introduction

Correlations between physical properties of galaxies are helpful to extract key variables that may account for the variety in galaxy properties (Lagos et al. 2016). The most fundamental of these correlations (called scaling relations) can effectively simplify the diversity of galaxy properties, however, they do not appear in any model or numerical simulation of galaxy formation and evolution. They are used only a posteriori to check the goodness of our models in reproducing the observed correlations, and galaxy formation models still have problems with many basic relations of galaxies, such as those including colours, metallicities, shapes, angular momentum, star formation rate, and initial mass function (Dekel & Birnboim 2006; Fall & Romanowsky 2013; Dutton et al. 2011; Cappellari et al. 2012).

The scaling relations are always in 2D or 3D because they are most often identified through trial and error, more rarely from statistical tools such as principal component analysis (PCA) or

the multivariate analyses that are explicitly aimed at minimizing the number of variables needed to explain a large fraction of the variance in the p -dimensional (with $p \gg 3$) space of variables. Although the scaling relations provide valuable insight, ultimately they cannot by themselves distinguish between cause and effect, past and future (see e.g. Lagos et al. 2016). This may be due to several reasons, one being that the scaling relations often show variability within or across the line/plane, this scatter being often correlated to another variable; this shows the complexity of galaxy physics and evolution. Another reason can be that the scaling relations reveal some hidden (latent) variable that has not been or cannot be identified yet (Fraix-Burnet 2011).

Despite these problems, scaling relations provide very useful constraints for models; for example, the mass–metallicity relation (see e.g. Faber 1973) suggests the paths of chemical evolution and indicates the amount of inflow and outflow processes across cosmic epochs, while the black hole–bulge mass relation (see e.g. Magorrian et al. 1998) suggests the co-evolution

of these two structures, and the luminosity–size relation (see e.g. [Kormendy 1977](#)) constrains the epoch and location of star formation in the galaxy halo. Other relations, for example the velocity–luminosity relations (the Faber–Jackson [Faber & Jackson 1976](#), the Tully–Fisher relation [Tully & Fisher 1977](#)), or the fundamental plane ([Djorgovski & Davis 1987](#); [Dressler et al. 1987](#)), indicate the main physical laws driving the formation of galaxies and what are the main causes of the properties observed in the correlations. All this information is extremely important for tuning models and simulations that today reproduce quite well dark matter properties, but fail in reproducing the whole range of observable properties of stellar structures, probably due to the complexity of baryon physics at local scales.

In this work we want to explore an alternative approach that we hope can enable further progress in this research area. This approach can be summarised with the following question: what can we understand about the evolutionary paths of galaxies if we start from the observed scaling relations? In other words, what can be deduced when a statistical multivariate analysis is applied to a set of galaxy variables? It must be realised that the multivariate analyses, especially phylogenetic approaches, do reveal the equivalent of “scaling relations”, but in a space whose dimensions are given by the number of variables. This is much harder to visualise with plots representing real data and this is why many visualisation tools exist (the typical cladistic tree being one of them) but they are still difficult to compare to physical models directly. The phylogenetic approach can give a kind of chronology provided that the trees can be rooted (oriented) thanks to some ancestrality hypothesis. This means that once we have at our disposal data for galaxies at different cosmic epochs, we could use the phylogenetic approach to identify the main roots (primitive kinds of galaxies) of evolution and derive useful constraints for our models. Different formation histories might in fact result in different types of correlations between galaxy variables.

Unfortunately galaxy variables across cosmic epochs are often difficult to obtain and are subject to numerous sources of bias. Good and abundant measurements are available today only for nearby galaxies, since high redshift objects are still sparsely observed and at the limit of present day telescopes. However, in recent years thanks to the development of galaxy surveys, such as WINGS ([Fasano et al. 2006](#)), the Sloan Digital Sky Survey (SDSS; [Abazajian et al. 2003](#)), SAURON ([Bacon et al. 2001](#)) and many others, it was possible to collect a lot of galaxy measurements for hundreds and thousands of galaxies that can be used for our statistical investigation.

This work, based on the robust statistical sample of nearby galaxies in WINGS clusters, tries to use multivariate statistical analysis to check the possibility of understanding the preferential paths of galaxy evolution hidden within the complex interrelations amongst the structural variables. At this moment, the sample involves galaxies in the redshift range 0.04–0.07, at a typical radial comoving distance of ~ 250 Mpc. In the near future, works with data for high redshift galaxies might complete the construction of the main paths of galaxy evolution.

The paper is organised as follow. In Sect. 2 we describe the WINGS data sample used in this work and we give a short review of the galaxy variables used in our analysis. Section 3 introduces the multivariate analysis approach, in particular the maximum parsimony method (Sect. 3.1). We discuss the selection of the variables used in the phylogenetic analysis in Sect. 3.4 and we present our results in Sect. 4 discussing three particular cases: the full sample (Sect. 4.1), the cluster sample (Sect. 4.2) and the field galaxy sample (Sect. 4.3). The discussion follows in Sect. 5 and our conclusions in Sect. 6.

2. Data

The data used in this paper come from the WINGS database ([Moretti et al. 2014](#))¹.

2.1. The Wide-field Nearby Galaxy-clusters Survey sample

WINGS is a long term research project especially designed to provide a robust characterisation of the photometric and spectroscopic properties of galaxies in nearby clusters ($0.04 < z < 0.07$; [Fasano et al. 2006](#)). The core of the survey WINGS-OPT ([Varela et al. 2009](#)) is the database of optical *B* and *V* images of 77 clusters, obtained during dark time with the Wide Field Camera (WFC, $34' \times 34'$) mounted at the corrected $f/3.29$ prime focus of the INT-2.5 m telescope in La Palma (Canary Islands, Spain) and with the Wide Field Imager (WFI, $34' \times 33'$) mounted at the $f/8$ Cassegrain focus of the MPG/ESO-2.2 m telescope in La Silla (Chile). The clusters have been analysed with SExtractor ([Bertin & Arnouts 1996](#)). The optical photometric catalogues are 90% complete at $V \sim 21.7$, which translates to $M_V^* + 6$ at the mean redshift of the survey ([Varela et al. 2009](#)). The WINGS-OPT database includes, respectively, 393013 galaxies in the *V* band and 391983 in the *B* band.

The WINGS database has a spectroscopic follow-up WINGS-SPE for a subsample of 48 clusters (26 in the northern and 22 in the southern hemisphere) with the Multi-object Fibre spectrograph at the *William Herschel* Telescope WYFFOS at WHT (λ range = $3800 \div 7000$ Å, resolution $FWHM = 3$ Å) and the Two Degree Field spectrograph at the Anglo-Australian Telescope 2dF at AAT (λ range = $3600 \div 8000$ Å, resolution $FWHM = 6$ Å). From these spectra we derived the redshift of 6132 galaxies ([Cava et al. 2009](#)). A subsample of this (5299 galaxies) has been analysed with spectro-photometric techniques, deriving the star formation rate sfr^* at different epochs ($sfr1$, $sfr2$, $sfr3$, $sfr4$), the stellar masses M^* and ages, the internal extinction A_V , and the equivalent widths of the main absorption features (see for all details [Fritz et al. 2007, 2011](#)).

A full re-analysis of the WINGS spectra, aimed at measuring the emission line properties, has also been performed by [Marziani et al. \(2017\)](#). These authors considered 46 clusters of the WINGS-SPE survey, and performed a population synthesis from spectra that were not considered in [Fritz et al. \(2011\)](#): spectro-photometric data were available for 5859 sources in the fields of 46 clusters. Of these, 3514 spectra were of cluster galaxies. All cluster spectra were then joined to form a stacked sample with 3514 spectra, while 2344 spectra of non-cluster sources were considered as the field sample.

2.2. Adopted sample

From the WINGS database, a sample of 4100 galaxies has been extracted for this work. Distances were determined from their redshifts. We determined 2624 galaxies to be members of 28 galaxy clusters (Fig. 1) with redshifts around 0.05–0.06. The remaining 1476 galaxies are background galaxies with slightly larger redshifts around 0.15 (Fig. 2). They are considered as field galaxies. This field sample is heterogeneous, since it is represented by galaxies in the field of our clusters. Binary galaxies or small groups might be present. The possible presence of subgroups has been analysed by [Ramella et al. \(2007\)](#). Cluster

¹ The WINGS website provides all the details of this project at <http://www.pd.astro.it/index.php/en/nearby-clusters-wings.html>

Fig. 1. Histograms of cluster and field sample galaxies in the clusters.

Fig. 2. Gaussian kernel density distribution of the seven variables used in the clustering and maximum parsimony computation (Table 1), plus the absolute V magnitude, the morphology T and the redshift, for the cluster (solid line) and field (dotted line) samples. The scale on the y-axis is given by the integral of the kernel density estimation, which is normalised to one.

background galaxies have been eliminated on the basis of the colour-magnitude diagrams (see [Valentinuzzi et al. 2011](#)) and on the results obtained by [Ramella et al. \(2007\)](#). In general, galaxies redder than $B - V = 1.4$ that do not belong to background clusters have been kept. The reader should keep in mind that this

Table 1. Variables available for the WINGS sample.

$B - V$	$B - V$ colour
$\log(R_e)$	Logarithm of effective radius [kpc]
n_V	Sersic index
$\langle \mu \rangle_e$	Mean effective surface brightness
H_β	Equivalent width of H_β
D_{4000}	Balmer discontinuity index
$\log(M^*)$	Logarithm of stellar mass
M_V	Absolute V magnitude
T	Morphological type
ϵ	Ellipticity
Δ_{CC}	Projected distance to cluster centre
$sfr1$	Star formation rate in $M_\odot \text{ yr}^{-1}$: stellar mass produced within the age bin $0-2 \times 10^7 \text{ yr}$
$sfr2$	Idem of $sfr1$ for the bin $2 \times 10^7-6 \times 10^8 \text{ yr}$
$sfr3$	Idem of $sfr1$ for the bin $6 \times 10^8-5.6 \times 10^9 \text{ yr}$
$sfr4$	Idem of $sfr1$ for the bin $5.6 \times 10^9-17.8 \times 10^9 \text{ yr}$
$M1/M4$	Stellar production ratio: $(sfr1 \times 2 \times 10^7) / (sfr4 \times (17.8 \times 10^9 - 5.6 \times 10^9))$
$W(H\alpha)$	Rest-frame $W(H\alpha)$
$[NII]/H\alpha$	Intensity ratio $[NII]\lambda 6584/H\alpha$

Notes. Only the seven first parameters (top panel) have been used for the clustering and maximum parsimony analysis.

fact could partly affect our conclusions when the cluster and field samples are compared.

The Galaxy Surface PHOTometry (GASPHOT) tool used in WINGS to derive the radius, surface brightness, and ellipticity takes into account the convolution with the point spread function (PSF). Hence, the model profiles include the PSF measured on nearby stars. In addition the smallest galaxies always cover a surface of 200 pixel squares, so the minimum radius is about eight pixels while the typical stellar PSF of the survey is around three pixels. The light profiles are clearly seen even if the pixel scale is different for field and cluster objects. All the parameters measured for cluster and field objects are thus equally well derived, alleviating any potential biases in the comparison of the properties between cluster and field objects.

In this paper, we consider 18 galaxy properties (see Table 1). More variables are available, but some of them have a large fraction of unavailable values. The set of seven parameters used for the cladistic analysis itself (see Sect. 3.4) is reported in the top part of Table 1. They are: the total $B - V$ galaxy colour corrected for galactic extinction and K -correction; the logarithm of the circularised effective radius $\log(R_e)$ in kiloparsec units; the Sersic index n_V in the V band obtained by GASPHOT ([D’Onofrio et al. 2014](#)) from the fit of the seeing-convolved major and minor axes surface brightness profiles; the mean effective surface brightness in the V band; the equivalent width (EW) of the H_β line visible in the spectrum, following [Fritz et al. \(2011\)](#); the Balmer discontinuity at $4000 \text{ \AA}$ D_{4000} ; the logarithm of the stellar galaxy mass $\log(M^*)$ derived from the spectral energy distribution (SED) fitting ([Fritz et al. 2007](#)). The distributions of these seven parameters for the cluster and field samples are essentially the same except for $B - V$, n_V and $\log(M^*)$, field galaxies being somewhat more massive, bluer, and with a smaller Sersic index (Fig. 2). This implies that any difference in clustering results between the two populations cannot be explained simply by some monovariate differences.

The other variables are used to describe the characteristics of the groups and their mutual relationships. These variables

include: the total galaxy luminosity M_V , the morphological types derived with the MORPHOT tool by [Fasano et al. \(2012\)](#), the galaxy mean ellipticity derived from the GASPHOT analysis ([D’Onofrio et al. 2014](#)), the projected distance from the cluster centre brightest cluster galaxy, the star formation rates ($sfr1$, $sfr2$, $sfr3$ and $sfr4$) computed in four age ranges ($0-2 \times 10^7$, $2 \times 10^7-6 \times 10^8$, $6 \times 10^8-5.6 \times 10^9$ and $5.6 \times 10^9-17.8 \times 10^9$ yr), the ratio between the stellar mass formed at the latest epoch to the mass formed at the earliest epoch (that we will call the stellar production ratio $M1/M4 = (sfr1 \times 2 \times 10^7)/(sfr4 \times (17.8 \times 10^9 - 5.6 \times 10^9))$), the equivalent width $W(H\alpha)$, and the intensity ratio $[NII]\lambda 6584/H\alpha^2$. With the exception of the stellar mass M^* all variables come from direct measurements made on images and spectra. They are not model dependent.

3. Analyses

We first introduce here the maximum parsimony technique used in this work.

3.1. Maximum parsimony analysis

Phylogenetic approaches seek to establish relationships between objects or classes of objects. They are looking for the simplest evolutionary scenario to transform any object of the sample into any other one. The result is a graphical representation (most often a tree) on which objects are gathered according to their evolutionary closeness. The phylogenetic technique can be distance-based if pairwise (dis)similarity matrices are computed from the parameters, or character-based if the parameters are used directly. These parameters are called characters if they can be given discrete values that represent evolutionary steps. In the ideal case, innovations occurring over the course of evolution are transmitted to all the descendants, allowing the reconstruction of the entire diversification history of the sample. The characters can be discrete, continuous, or categorical (nominal, qualitative) in the case of genetic data. In astrophysics, most variables are quantitative and continuous.

The maximum parsimony (or cladistics, hereafter MP, [Hennig 1965](#); [Felsenstein 1984](#)) technique is character-based and the most general phylogenetic tool. Because it looks for the simplest path connecting the objects under study, it has some similarity with the minimum spanning tree technique that has been extensively used to determine galaxy clusters in order to map the spatial distribution of baryonic matter ([Barrow et al. 1985](#); [Bhavsar & Splinter 1996](#)). However, in the MP technique, internal nodes are introduced, creating a much larger variety of tree topologies at the expense of a much heavier computation.

Amongst all the possible arrangements of the objects on a tree, MP selects the one that has the lowest total number of step changes (the score). A step change is the absolute difference between the values of a parameter (character) between two nodes of a branch (edge) of the tree. By minimising the score, the algorithm selects the simplest evolutionary path that links all the objects. Mathematically, the score of a tree corresponds, after labelling of the internal nodes, to the minimum number of edges (u, v) , with $c(u) \neq c(v)$, $c(u)$ being the character state at node u . A real value $f(u)$ is associated to each internal node. The score s of a tree equals the sum over all edges of the absolute difference between those values:

$$s = \sum_{e=(u,v) \in E} |f(u) - f(v)|. \quad (1)$$

This can be directly extended to continuous characters or values.

As a consequence, two objects that are close on the tree share similar evolutionary histories. In this way, we can define evolutionary “families” that are called “clades” in biology but that we will simply call groups. This approach has been explained in detail in many papers for astrophysics (e.g. [Fraix-Burnet et al. 2006a,b, 2015](#); [Fraix-Burnet 2016](#); [Holt et al. 2018](#)), and the simplest illustration of the power of the phylogenetic approach in astrophysics is the reconstruction of the stellar evolutionary paths and of the families of stars born together in open clusters, a problem known as “chemical tagging” ([Blanco-Cuaresma & Fraix-Burnet 2018](#)).

In this paper, the parameter values have been discretised into 32 bins, and the MP computations were performed with PAUP 4.0 ([Swofford 2003](#)). There are generally several to many equally most parsimonious trees, so we made a consensus tree that maintains the appearance of internal nodes in at least 70% of them. We aim at what is called a “resolved” tree (or binary tree) where each internal node has only three branches. In this case, the evolutionary scenario depicted by the tree is more informative (it is said that the phylogenetic signal is strong). This is the criterion we used to select the best subset of parameters (see Sect. 3.4). The reliability of the result was then estimated through several methods:

1. Other subsets of parameters: in theory the best indicator of robustness is a bootstrap approach. Since the computation time is too large, we simply ran several analyses with some parameters removed, and compared tree structures and the evolution of the parameters along the trees.
2. We performed a neighbour joining tree estimation ([Saitou & Nei 1987](#); [Gascuel & Steel 2006](#); [Fraix-Burnet et al. 2015](#); [Jofre et al. 2017](#)) using both the Euclidean and the Manhattan (L1 norm) distances. This is a different but also quite popular phylogenetic technique, based on distances between objects. This is a bottom-up hierarchical clustering method that minimises tree length, according to a criterion that can be viewed as a balanced minimum evolution ([Gascuel & Steel 2006](#)).
3. Using the median values of the parameters for each group defined from the MP analysis, we ran a branch-and-bound algorithm ([Hendy & Penny 1982](#)) that thoroughly explores all the possible tree topologies, and we selected the most parsimonious one. This algorithm is much more efficient in finding the optimal tree than the heuristic one used for the three samples, but it can only be used on a small number of objects.

The fact that all these analyses agree is a good indicator that a phylogenetic signal is present in the set of parameters used.

3.2. Pre-clustering with the hierarchical clustering technique

Since the computing time to explore the tree space is prohibitive beyond 800 objects or so, we decided to apply a pre-clustering technique. The idea is that we are not aiming at obtaining a genealogy of galaxies, which is not feasible, but a phylogeny, in other words a diagram depicting the evolutionary relationships between classes of galaxies, these classes being unknown at this stage. Each galaxy of our sample is first considered as a potential exemplar representing its own class, but similar objects are expected to be redundant. Hence, we can try to group them together into what we will call sub-clusters. A class is thus made

² The values for the emission lines $H\alpha$ and $[NII]$ are not available for all galaxies and have been simply disregarded in the corresponding plots.

Fig. 3. Three possible views of the phylogenetic tree for the full sample of galaxies. *Left:* tree with the group definition. Only groups with more than 50 galaxies are considered for this tree in this paper. The tree has been rooted with the group (at the top) with low values of $B - V$, n_V and $\log(M^*)$. *Middle:* unrooted tree. *Right:* tree with the proportion of cluster galaxies within each of the 300 pre-clusters indicated by the length of segments.

of one or more sub-clusters. We can adapt the number of sub-clusters to the computational constraint of the MP analysis and to the expected or desired number of classes. The advantage of this approach is that we can use very efficient methods of multivariate clustering with the pairwise distances (between each pair of galaxies) using the p -parameters. The most logical technique for our particular goal seems to be the hierarchical clustering approach.

Each sub-cluster is then characterised by the median values of the parameters computed from all the galaxies belonging to this sub-cluster. The pre-clustering is done independently on the three samples, full, cluster, and field, with the same number of pre-clusters each time. The MP analysis is performed on the sub-clusters, and the classes defined from the resulting tree gather together one or several sub-clusters.

We are now faced with the choice of the number of sub-clusters. For the MP analysis, one thousand is a maximum otherwise the computing time becomes prohibitive and the search for the most parsimonious tree may be less efficient. In any case, we do not expect that many classes of galaxies, at least for this still exploratory work, a few tens being already quite satisfactory. We considered that 300 pre-clusters is a good compromise. We also performed the same analyses with 100 and 200 pre-clusters to check the stability of the results on the full sample (see Appendix A). We cannot expect a perfect match since the minimised paths between different sets of pre-clusters cannot be identical. However, the global evolutionary scenario is preserved, while some groups might be displaced and/or split. The hierarchical clustering was performed with the function “hclust” of the package “stats” in R, with an euclidean distance and the method “ward.D2” as recommended in [Murtagh & Legendre \(2014\)](#).

3.3. Defining groups from a tree

Unlike partitioning techniques, clustering methods based on graphs, such as hierarchical clustering and phylogenetic tools,

require some choices in the definition of classes or groups since there are many possible sub-graphs. Nevertheless this reflects the fact that there are several possible classifications for the same data set depending on the purpose. This is well known in biology with the hierarchical organisation into species, genus, and family. The choices depend on the granularity needed for the study and are necessarily a compromise between the two extremes: there can be as many classes as galaxies implying that each galaxy is considered to be unique, or there can be only one class with all galaxies.

In our case, we first reduced the granularity by performing a pre-clustering through the hierarchical clustering technique (Sect. 3.2) to ease the visualisation of the relationships between different kinds of galaxies. The number of pre-clusters is constrained by the computation time of the subsequent MP analysis and is a good representation of the diversity of the sample.

The definition of the groups from the MP tree is firstly guided by the structure of the tree: a group can be defined when a bunch of branches starting from a node is clearly apart from other bunches of branches (see Fig. 3 for an illustration). Such groups may potentially define a kind of family or species sharing some properties resulting from the evolution. The goal here is not necessarily to establish a firm general classification but to find the relevant level of granularity that helps in understanding the diversification of galaxies.

We must point out that classes of galaxies could be defined by a partitioning or a hierarchical clustering technique (that is our pre-clusters). However, the MP analysis provides complementary information with the relationships among these classes. As a matter of fact, this is the first time that a hierarchical (clustering and phylogenetic) classification has been attempted on such a large sample.

3.4. Selection of the parameters

Ideally we would like to keep all available variables but there is no guarantee that a phylogenetic signal is present or can

be found. The variables must bear some reliable information on the evolutionary history of the galaxies. As a consequence, we should disregard variables with convergent and regressive behaviours, as well as redundancies, because they bring confusion regarding ancestry and can destroy the tree structure. Since it is not always easy to detect these “flaws” a priori, we need to try different subsets and keep the results that show the most parsimonious trees that are as resolved as possible. We always try to choose the subset with the largest number of variables to be more informative, more objective, and in better accordance with the complexity of galaxies and their evolution.

For the analysis itself, we did not use the star formation rates sfr^* at different epochs, which are derived, model-dependent variables, and the total galaxy luminosity M_V , which is related to $\log(R_e)$ and $\langle\mu\rangle_e$ by a linear formula. We also do not consider the de Vaucouleurs’ morphological type index T because it is a categorical value, and the distance to the cluster centre Δ_{CC} for the galaxies belonging to a cluster, because it is not an intrinsic descriptor of a galaxy. A first analysis included the ellipticity ϵ , which was found to be somewhat erratic on the tree as compared to the other parameters. We decided to disregard it since it is affected by orientation effects.

We end up with the seven parameters described in the top part of Table 1: $B - V$, $\log(R_e)$, n_V , $\langle\mu\rangle_e$, H_β , D_{4000} , and $\log(M^*)$. The disregarded variables are also reported in the bottom part of Table 1 and are used for the interpretation.

It should be clear that the selection of the parameters is not as arbitrary as it might look. Firstly there are solid arguments to disregard some the observables as explained above. Secondly, the robustness of the tree is a strong statistical argument. Finally, any multivariate statistical analysis is an objective and powerful way to observe the data space, but only the physical interpretation of the results is able to tell its real interest and relevance. Any obvious contradiction should lead us to reconsider the selection of the parameters as we did for the ellipticity.

The seven parameters that have been selected mainly from a statistical point of view constitute in fact a reasonable description of the physics related to the stellar content of galaxies: colour (metallicity and age of the stellar population), structure (radius and light distribution), star-forming activity, relative importance of old and young stars, and total stellar mass. We here stress that the selection of the parameters is based on purely statistical arguments, that is, the robustness of the tree we can obtain. There is no physical assumption here except for the fact that the “simplest” tree chosen in the MP analysis corresponds to the smoothest evolution of the seven parameters. This means that, as said at the beginning of this section, we consider that these parameters are tracers of the evolution of galaxies. For instance, the intensity of H_β in emission is a short-lived feature depending on stellar bursts, while its equivalent width (used in this paper) is a more stable feature of the dynamics of the galaxy. The tree obtained is an hypothesis for an evolutionary scheme and its interpretation provides physical justification or rebuttal (see an example in Fraix-Burnet et al. 2006a, for a physical inconsistency of the resulting tree due to burst indicators).

4. Results

We performed three identical analyses (same seven parameters listed in Sect. 3.4), pre-clustering, with 300 pre-clusters (Sect. 3.2), followed by the MP analysis (Sect. 3.1) separately on the full sample (4100 galaxies, Sect. 4.1), the cluster sample (2624, Sect. 4.2), and the field sample (1476, Sect. 4.3). Our goal

with the full sample was to look for possible separate evolutionary paths between cluster objects and field ones. In other words, are they evolutionary disconnected, do they share some common ancestor from which their evolution diverged, or did they evolve similarly but at a different pace due to their environment? Since we found that the latter is true, we decided to conduct analyses of the two sub-samples to look for more subtle differences in their evolution as separate populations and shed some light on the precise role of the environment in the statistical evolution of galaxies.

4.1. Full sample

The tree with the 300 hierarchical clusters is nearly perfectly resolved and is well structured (Fig. 3). On this figure, the tree is shown in two graphical representations: the one to the left has been rooted with the group (at the top) that has low values of $B - V$, n_V , and $\log(M^*)$, and the one to the right is unrooted and is the real outcome of the MP analysis. Rooting a tree is merely an aid to evolutionary interpretation.

It is obvious that the evolution of the properties along the tree, hence the evolutionary scenario that is depicted by the tree, depends on the rooting. The analysis is independent of this choice. Rooting a tree consists in choosing the group that is considered as the most ancestral and drawing the tree starting from this group. We have chosen a group that is blue, has a low Sersic index and a low mass. We show in Figs. A.5 and A.6 a different rooting, on group A7, which could be justified as having a moderate mass and as having formed stars with very early and very low star formation rates $sfr1$, $sfr2$, and $sfr3$. Clearly this new rooting does not yield a satisfactory evolutionary scenario.

The groups are defined from the structure of the tree guided by bunches of branches that are clearly apart from other bunches of branches (Sect. 3.3). As we can see, more (smaller) groups (each corresponding to a branch in a bunch of branches) could have been defined but this would complicate the discussion in the paper. We recall that each tip (leaf) of the tree is a pre-cluster (see Sect. 3.2) of one or several galaxies. For reasons of simplicity in the discussion and interpretation in this paper, we only consider fifteen groups, which have more than 50 galaxies (A1–A15, Fig. 3), thus leaving only 480 (11.7%) galaxies aside.

The tree in Fig. 3 is quite regular, with A4 and A5 together making a significant substructure. The distribution of the galaxies in the groups (Table 2) is roughly balanced, with the two smallest groups gathering 64 and 94 galaxies (A14 and A13 respectively), and the two largest ones gathering 477 and 434 objects (A5 and A2 respectively). We here note that parts of groups A4 and A11 are displaced towards the top on the tree with 200 pre-clusters (Appendix A) suggesting some uncertainties on their precise placement.

The distribution of cluster and field galaxies is quite different in the groups (Table 2): eight groups (A3, A5, A7, A9, A11, A12, A13, and A15) are composed mainly or nearly exclusively of cluster galaxies (from 73% to 96% to be compared to 62% for the full sample), five groups (A1, A2, A6, A8, and A10) are dominated by field galaxies (since they have only 24% to 41% of cluster galaxies), and two (A4 and A14) have an equal share (48% and 52% of cluster galaxies). There seems to be an increase in the percentage of cluster galaxies from A1 to A15, with large fluctuations, which can be visualised from the proportion of cluster galaxies within each of the 300 pre-clusters (Fig. 3). These large fluctuations indicate that neither the hierarchical clustering nor the MP analysis are able to separate clearly the two galaxy populations despite some very few exceptions.

Table 2. Distribution of galaxies in the different groups from the analysis with the full sample, with numbers and proportions of cluster and field galaxies for each group.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	Total
All galaxies	208	434	344	297	477	231	162	128	182	172	139	362	94	64	326	3620
Cluster galaxies	77	116	265	144	405	55	143	41	146	70	134	299	86	33	239	2253
Field galaxies	131	318	79	153	72	176	19	87	36	102	5	63	8	31	87	1367
% cluster galaxies	37	27	77	48	85	24	88	32	80	41	96	83	91	52	73	62
Composition of the remaining 21 small groups:																
	6	13	14	15	15	16	17	18	19	20	22	22	23	25	25	27
	32	34	37	40	40											

Notes. The table also lists the numbers of galaxies in the 21 small groups not considered in this paper.

Fig. 4. Violin plots for fifteen variables and the groups A[1–15] of the full sample as defined on Fig. 3. The seven first parameters only (in red) were used for the analysis. The horizontal bar is the median value, and each “violin” symmetrically depicts the distribution of the variable.

The tree does not show separate evolutionary paths for cluster and field galaxies. However, assuming that the rooted tree reveals an evolutionary scenario, the increasing importance of cluster galaxies could be interpreted as them being somewhat more diversified on average than field galaxies. By diversification we mean that the objects become more different from a common ancestor supposedly situated at the top of the tree. Fraix-Burnet et al. (2010) have shown that “diversification” is preferable to “evolution” since the latter is difficult to define and measure in a multi-parameter space for a whole population. In the present case, despite the fact that field galaxies are more massive on average (Fig. 2), they are not the most “evolved” objects, probably because they are also still blue and concentrated. Member and field galaxies evolved but underwent different transforming-process histories: since cluster galaxies appear more diversified, one obvious explanation is their environment, which makes them suffer from more frequent and/or violent events.

The properties of the groups are represented as violin plots in Fig. 4. This representation gives the median values and the distribution for each variable as a function of the group index in the order given by the tree on Fig. 3. This sequence provides the changes of the variables along the diversification process and should not be taken literally as the evolutionary steps followed by individual galaxies.

The colour $B - V$ increases sharply up to group G5 and then becomes roughly constant except for a drop at Group A14. The effective radius $\log(R_e)$ wiggles with no clear evolutionary

trend. The Sersic index n_v , also used to root the tree, increases regularly. The effective surface brightness $\langle \mu \rangle_e$ decreases globally despite some exceptions (A4 and A12 for instance). H_β and D_{4000} have a very similar behaviour as $B - V$, but the D_{4000} value of A7 is remarkably higher than for all the other groups. $\log(M^*)$ seems to be lower from A1 to A5 and higher afterwards, but there are many exceptions, while M_v behaves more or less in the opposite manner. ϵ does not show much significant difference between the groups since the intra-group variance is large. Groups A1, A2 and A3 are spiral-dominated groups with the morphology variable T higher than for all the other ones.

The different star formation rates are shown in Fig. 4 as well: $sfr1$ and $sfr2$ look quite different in groups A1 to A8 and more similar in the subsequent groups. On the contrary, $sfr3$ and $sfr4$ seem globally lower in the first part (up to group A7) with two exceptions: A6 and A14. Finally, the stellar production ratio $M1/M4$ is lower from A7 to A15, except for A14 and possibly A10.

The properties as given by the violin plots explain the structure of the tree and allow a physical understanding of the groups. The gross picture of galaxy evolution is correctly recovered. Galaxies become redder and more massive, with ageing stars; the most diversified galaxies formed stars earlier in the past and they tend to reach an early-type morphology. We see an increase of the effective surface brightness, but no clear evolution of the effective radius. There are some peculiar groups that require further investigation (Sect. 5), such as A7, with its high D_{4000} , or A14, with its astonishingly low $B - V$, low mass, low D_{4000} , and

Fig. 5. Tree from the analysis with the cluster sample. Only groups with more than 50 galaxies are identified. The tree on the left has been rooted with the group (at the top) that has low values of $B - V$, n_V and $\log(M^*)$. To the right we show the same tree but unrooted.

low star formation rates at earlier times ($sfr3$ and $sfr4$) suggesting they could be rejuvenated galaxies.

4.2. Cluster galaxies

The tree resulting from the analyses of the 2624 cluster galaxies (Fig. 5) and rooted as for the full sample (using low values for $B - V$, n_V , and $\log(M^*)$) is perfectly resolved and nicely structured. We only consider the twelve groups that have more than 50 galaxies (C1–C12, Fig. 5), leaving 336 (12.8%) galaxies aside.

Three groups dominate (C5, C11, and C12, Table 3), and three are significantly smaller (C3, C6, and C10). On the tree, after C5, there is a split into two clear ensembles: C6 to C9 on one side, and C10 to C12 on the other. This is important to keep in mind when interpreting the evolution of the variables and we must not be deceived by the labelling: after C5, the evolution of the variables splits into two branches, one towards C6, the other one towards C10.

The violin plots (Fig. 6) show a similar evolution of all the variables along the tree as for the full sample tree, except that there is a clearer increase of mass $\log(M^*)$ (and a decrease of M_V), as well of $sfr3$ and $sfr4$. There are also three clear regimes for the morphology T : spirals for C1 and C2, S0s for C3 to C7 and C9, and ellipticals for C10 to C12. We recall that the latter three form a specific ensemble on the tree. Departing from this scheme, C8 is composed of spirals.

It is tempting to associate C6 to A7 (being composed of 88% of cluster galaxies) considering their peculiar high D_{4000} value. This is corroborated by low $sfr1$, $sfr2$, and $M1/M4$ in both cases. However C6 is made of 59 galaxies while A7 has 143 cluster galaxies.

The median distances from the centres of the clusters (Fig. 6, last panel at the bottom right) show that the groups C3, C7, and the ensemble C10, C11, and C12, are the most central ones even though the differences are not highly significant.

Despite the global similarities in the evolution of the variables, which reflects the approximate picture of galaxy evolution, the tree for the cluster galaxies is characterised by two

distinct evolutionary paths, one from C6 to C9, and one from C10 to C12 (Fig. 5). The evolution of some variables is clearly distinct along these two branches. The value of $B - V$ decreases in the first branch from C6 to C9 and increases in the second one from C10 to C12. The Sersic index n_V is very low in the first branch, but increases strongly in the second one. D_{4000} decreases in the first branch and increases in the second one. In the second branch, $\log(M^*)$ increases strongly, ϵ is lower, and T is much lower. Interestingly, $sfr1$ and $sfr2$ increase in both branches, while $sfr3$ and $sfr4$ decrease in the first branch and increase in the second one. These different star formation histories are clearly visible in the stellar production ratio $M1/M4$, which increases in the first branch and decreases in the second one. Finally, the second branch contains more central galaxies (lower Δ_{CC}). All this might suggest that the first ensemble (C6 to C9) are galaxies that increase their stellar mass by forming stars, while the second ensemble (C10 to C12) are galaxies that increase their stellar mass by accreting old stars probably formed in other galaxies, that is, by dry mergers.

4.3. Field sample

The tree (Fig. 8) obtained with the 1476 galaxies of the field sample is fully resolved and very regular. The rooting is the same as for the other trees in this paper, with low $B - V$, n_V , and $\log(M^*)$. To keep a similar number of groups as for the two other samples, we identify fifteen groups with more than 23 galaxies, gathering 1158 and thus leaving 318 (22.5%) galaxies aside (Fig. 8). We note that group F6 with 31 galaxies is one of the 300 pre-clusters and thus occupies a single branch on the tree. Four groups dominate (F3, F7, F13, F15, Table 4) with 101 to 219 galaxies, and F12 is the smallest with only 24 galaxies.

The violin plots (Fig. 7) show that $B - V$ is somewhat more chaotic than for the full and the cluster samples, but it is higher on average on the second part of the tree, hence suggesting an increase. The Sersic index n_V seems to increase more regularly than for the cluster sample, while the increase of $\log(M^*)$ is less obvious except that the two first groups have a much lower mass. For the other variables, the notable differences with Figs. 4 and 6 are $\langle \mu \rangle_e$ that drops at F11 and then increases, and H_β that is more irregular.

There are two groups with a high D_{4000} (F9 and F12) but they do not share other similar properties with A7 and C6, except maybe for F12 that has a low stellar production ratio $M1/M4$. These two groups, F9 and F12, are small but have more galaxies (respectively 33 and 24) than A7 (19 field galaxies).

The global picture of galaxy evolution is recovered as in the two previous cases, but it seems that the evolution of the variables looks slightly less regular than for the cluster galaxies or for the full sample. This implies a difficulty in finding a smooth evolutionary path between the pre-clusters, very probably because the distribution of field galaxies in the parameter space is too sparse. This can be explained in two ways: (i) the field sample is too small (its size is nearly half that of the cluster sample), so that the sampling performed by the hierarchical clustering in the seven-parameter space is too crude so that any evolutionary path would appear less regular. However, the size of the field sample is consequent and of the same order of magnitude as the cluster one. Or (ii) the population of field galaxies is intrinsically less homogeneous so that they occupy a wider and/or sparser region of the parameter space. The first explanation is an observational bias that could possibly be corrected by more data. The second explanation has a physical consequence: there are missing links in our sample, that is, galaxies that would help

Table 3. Distribution of galaxies in the different groups from the analysis with the cluster sample.[illegible]

Notes. The table also lists the numbers of galaxies in the 27 small groups not considered in this paper.

Table 4. Distribution of galaxies in the different groups from the analysis with the field sample.[illegible]

Notes. The table also lists the numbers of galaxies in the 51 small groups not considered in this paper.

Fig. 6. Violin plots for sixteen variables and the groups C[1–12] for the cluster sample. The projected distance Δ_{CC} of a galaxy to the centre of a cluster is also shown.

reconstruct the evolutionary scenario. Unfortunately, the missing links might not be observable at all, which would imply that the field galaxies evolved largely independently and lost most of the traces of their origin and their relationships to other galaxies. Nevertheless there is a significant difference between the two populations: the field sample has on average bluer, lower Sersic index, and more massive galaxies (Fig. 2), while we assumed that ancestral galaxies were similar but less massive. This suggests that field galaxies evolved more in mass than cluster galaxies. We conclude that our field sample is more heterogeneous than the cluster one, favouring an intrinsic physical origin without entirely ruling out a possible observational bias.

5. Discussion

We have proposed a phylogenetic analysis of a sample of galaxies, and obtained positive results from a methodological point of view. We must now discuss on the relevance of this approach for the study of the evolution of galaxies by first looking at the tree structures and the quality of the selected parameters for a phylogenetic reconstruction.

We stress that the discussion below is based on the rooting of the trees according to increasing $B - V$, n_V , and $\log(M^*)$. This choice is based on the a priori assumption that “primordial”

galaxies could be blue (because of young stars), with a small n_V and not very massive. This certainly can be disputed (see Appendix A) especially because our sample of modest redshift very probably does not contain any galaxies resembling primordial ones, but only objects with already a long and complex diversification history. However, we have used the luminosity-weighted age computed by Fritz et al. (2011) and notice that the low mass or low n_V correspond to a low age. These types of galaxies are present with higher prevalence amongst the C1 and F1. Hydrodynamical simulations seem to confirm that galaxies in the hierarchical scheme start with these properties (Vogelsberger et al. 2014; Genel et al. 2014; Snyder et al. 2015).

5.1. Influence of the parameters on the classification

In a multivariate analysis, it is always difficult to understand the origin of the clustering and the influence of the parameters. It is tempting to attribute the results to only a very few, if not one, of the parameters. However, we can now check statistically whether some of them are dominant, correlated, or if there exists some hidden variable explaining the clustering properties and the tree structures.

A first approach is to look for correlations between the seven parameters used in both the hierarchical pre-clustering and the

Fig. 7. Violin plots for sixteen variables and the groups F[1–15] for the field sample.

Fig. 8. Tree for the field sample. Only groups with more than 23 galaxies are identified. The tree on the left has been rooted with the group (at the top) that has low values of $B - V$, n_v , and $\log(M^*)$. Group F6 is a pre-cluster and occupies a single branch of the tree indicated by the tick mark. To the right we show the same tree but unrooted.

MP analyses. The Pearson correlation coefficient is always lower than 0.5 for the full sample: 0.57 ($\log(M^*)$ versus $\log(R_e)$) for the cluster sample and 0.67 ($\langle\mu\rangle_e$ versus $\log(R_e)$) for the field sample. We conclude that there is no significant correlation that could influence our classification results, and as a consequence no hidden variable that would significantly influence several of the parameters.

This can be further checked with a PCA. As shown in Fig. B.1 (top left diagram), there is no really dominant eigenvector and the first four principal components (PCs) explain only 70% of the total variance, showing that the seven parameters used are non-redundant and required to cover the diversity of the galaxies of our samples. This is true for our three samples (full, cluster, and field). The three other diagrams in Fig. B.1 show the projection of the data on the scatter plot drawn by the first two PCs, together with the importance (loadings) of each parameter

in these two components. All parameters are approximately of equal amplitude. The main difference between the three samples is that $\log(R_e)$ is close to $\langle\mu\rangle_e$ for the field sample, while it is between the latter and the five remaining parameters. The conclusion of the PCA is that there is no parameter or hidden variable that dominates the variance of the samples. In other words, the distribution of our samples in the 7D-parameter space is not very distorted.

However, the variance, as analysed by PCA, is not necessarily related to the clustering properties of the data set (e.g. Chang 1983). Once the classification is established, the correlation of the parameters with respect to the class labels can be computed. This is shown in Fig. B.2 for the three samples with the pre-clusters and the MP groups. For the former classification (dashed lines), the three samples are similar, with $B - V$, n_v , and $\langle\mu\rangle_e$ slightly dominating (with correlation coefficients between about 0.4 and 0.6), while the mass $\log(M^*)$ is more important for the cluster sample (coefficient of about 0.7). Still, the correlations are not very high and are comparable for all parameters, with some noticeable differences between the samples for $\log(R_e)$ and $\log(M^*)$. Therefore, the hierarchical pre-clustering is driven more or less equally by the seven parameters.

The MP groups show more significant differences depending on the sample (solid lines in Fig. B.2). The full and field samples are very similar except for $\log(R_e)$, which has a correlation coefficient near zero for the field sample. Apart from $B - V$, the cluster sample always shows significantly different correlation coefficients, being again dominated by $\log(M^*)$. Globally the coefficients remain mostly below 0.65 for the three samples, except for n_v in the field and full samples, which reaches 0.8.

Clearly, the effective radius does not play a role in the classification for both the field and cluster samples, but the Sersic index is of great influence for the field sample. Since this difference is not seen with the pre-clusters, we must conclude that it is the evolutionary nature of the MP analysis that makes it visible. Colour and mass are also important, the latter having a higher correlation coefficient for the cluster sample.

Could the influence of n_v in the field sample only be due to a selection bias from the redshift? We do not think so because (i) the correlation coefficient between the groups and the redshift is very low (0.12), (ii) the correlation between redshift and n_v is low also (0.10), and (iii) the changes of the redshift with the

groups (Figs. 6 and 10) do not show significant variation along the tree except for the low redshift of group F2. This group is relatively redder for its position on the tree, but is remarkable for its mass, which is the lowest of all the groups with values on the same order as group C1. In addition, since the groups have been identified in the seven-dimensional space of the parameters, it is difficult to see how a Malmquist-type bias could affect only one parameter.

There are some specific quantities that estimate which parameters explain the cladistic tree. One of these indices (rescaled consistency index, Farris 1989) synthesises the smoothness of the evolutionary behaviour of the parameters on the tree. It is a different indicator as compared to the ones described above since it is computed on the median values of the parameters within each pre-cluster and, more importantly, it is related to the evolution of the parameters along the tree rather than to the classification established from it. This index is globally lower for the field sample, confirming the apparent higher wiggling of the parameters for this sample on the violin plots (Sect. 4.3). The “best” parameters are $B - V$, D_{4000} , $\log(R_e)$, and n_V , depending on the sample, $\log(M^*)$ being always the worst, but the differences are not very important, which shows that the entire tree structures are supported somewhat equally by the seven parameters. Interestingly, for the cluster tree, $\log(R_e)$ is one of the two most influential parameters for the structure of the tree, but does not play a role in the classification based on the tree (see above).

The conclusion of this section is that the seven parameters used for the pre-clustering and MP analyses are not significantly correlated between each other but are probably not independent. The classification is not influenced by any particular properties, but there is a clear difference between the cluster and field samples concerning the main drivers of the MP classification: even if colour and mass slightly dominate in both samples, the radius $\log(R_e)$ and the Sersic index n_V do not play the same role in the origin of the groups for the two samples. However, these two structural descriptors and the stellar content seem to be the main drivers of the tree structures, and hence the evolutionary scenario, the mass being not important. The existence of this difference is a clear influence of the environment, which in return appears to have induced a loss of the evolutionary history of the structures of galaxies.

5.2. Evolution along the trees

The evolution of the variable medians for each group along the tree for the cluster and field samples is shown in Figs. 9 and 10. The evolution scale on the abscissa is simply the index of the group and thus is only indicative of the relative evolutionary stages for each tree, without any match between the two sets of groups.

As already concluded from the violin plots, the global evolution of the variables is similar in both samples. In particular the starting and ending points of the evolution curves are relatively close, meaning that the least and the most diversified groups are similar in both samples. There are, however, differences in the details; the evolution for the field sample seems to wiggle more (that is, it is less monotonous), as in H_β , $\log(M^*)$, $sfr3$, and $sfr4$. In addition, the bumps in the curves appear sometimes opposite, as in $B - V$, M_V , $sfr1$, and $sfr2$.

The two paths of evolution can then be interpreted as covering similar ranges of evolutionary stages, but with possibly less homogeneity (or less continuity) in the field sample (Sect. 4.3). In other words, the fact that galaxies belong to a cluster tends to smooth out somewhat their current diversity (more continu-

Fig. 9. Changes of the medians of the groups along the tree for the cluster sample. Shaded regions show 1σ dispersion around the median.

Fig. 10. Same as Fig. 9 for the field sample.

ous diversity within clusters than globally) while at the same time creates an enhanced diversification from earlier more primitive galaxies (Sect. 4.1). This suggests that more frequent interactions tend to accelerate galaxy evolution and generate more similar objects.

This homogeneity difference is supported by the correlations between the parameters and the groups (Fig. B.2): since the correlation coefficients are higher for the field sample (Fig. B.2), the groups are more specific in the sense that they are better described by specific values of the seven parameters, meaning somehow less homogeneity.

The median values of $\log(M^*)$, $sfr1$, $sfr2$, $sfr3$, and $sfr4$ (Figs. 9 and 10) are systematically higher (and lower for M_V) for

the field sample. These galaxies are thus on average more massive, more luminous, and form more stars at all epochs. However, at the same time they are on average bluer (lower $B - V$) with a lower Sersic index, n_V , somewhat in contradiction with the general, expected evolutionary trend. Hence using only one property to characterise the evolutionary stage of a galaxy is misleading. This explains why the field galaxies are less present towards the bottom of the tree (Fig. 3). This also shows that the impact of the Malmquist bias is weak since in unsupervised clustering analyses only the ranges are important, and not the exact shape of the distribution. In summary, for a similar diversification stage, field galaxies seem more massive.

There are four parameters in common with Fraix-Burnet et al. (2010), who performed multivariate cluster and cladistic analyses of 56 low redshift galaxy clusters: $\langle\mu\rangle_e$, $\log(R_e)$, the mass, and the distance to cluster centre, even though our mass is the original gas mass from which all stars of the galaxy formed (in other words, our mass includes the mass of stars as well as some remnant gas mass) while theirs is the dynamical mass. The rooting is also different since we do not have the Mg index, but the two results agree on the global evolution for these parameters and that this evolution is not simply monotonically linear.

There are also five parameters in common with Fraix-Burnet et al. (2012), who performed a phylogenetic analysis of 424 early-type galaxies (ETGs): magnitude (V for us versus B), $\log(R_e)$, $\langle\mu\rangle_e$, H_β , and the mass (gas versus dynamical). We find only two slight differences: their mass increases while $\langle\mu\rangle_e$ is globally constant. However, the departures from these trends are large and this may not be very conclusive.

We must be aware that the evolution depicted by the tree depend somewhat on its rooting, and this has been chosen from a single parameter (metallicity) for the above studies, and three parameters (colour, Sersic index, and mass) in our case. If we follow the group interpretation found in Fig. 11 of Fraix-Burnet et al. (2012), we must conclude that the choices of the rooting for these three studies are satisfactory since no physical contradiction appears with our understanding of the broad picture of galaxy evolution.

5.3. Groups and clusters

We do not find any obvious correlation of the groups of the cluster sample with the galaxy clusters, indicating that our classification is “universal” for low redshift galaxies belonging to clusters. This is in line with our finding that cluster galaxies seem somewhat more homogeneous (Sect. 5.2). This is in agreement with Fraix-Burnet et al. (2010) and Valentinuzzi et al. (2011) and would seem to indicate that the formation histories of low redshift galaxies are not dramatically different from cluster to cluster. Indeed, the analysis of Fraix-Burnet et al. (2012) shows that groups are mainly explained by very diverse histories. This is supported by the barely significant dependence of the groups on the distance to the cluster centre, the intra-group variance being large (Figs. 4, 6, and 7). This is apparently not in agreement with Valentinuzzi et al. (2011), who find that some galaxy properties depend on the local environment, but this is probably because our groups are based on seven properties together and thus based on a more complete galaxy physics.

5.4. Evolution in stellar mass

The separate analyses show that the evolution in stellar mass of cluster galaxies is clearer and more regular than for the field sample (Figs. 9 and 10). The star formation rates ($sfr1 \dots 4$) are on

average lower for cluster galaxies and the masses are lower as well.

The stellar mass of a galaxy can in principle only increase through harassment or strong interaction. The mass $\log(M^*)$ increases in our three analyses, the trend being much clearer in the case of the cluster sample. However, even in this case, the growth is not monotonic, with groups C9 and C10 having a much lower mass than the trend. There are three possible explanations for this drop.

Firstly, since the tree represents the shortest path relating groups in a seven-dimensional space, it is possible that the projection on a one-dimensional axis (here $\log(M^*)$) creates an apparent drop. This would be the case for instance if one of the seven parameters showed a reversal in its evolution. Figure 6 does not reveal a unique culprit, none of the seven parameters being perfectly monotonic.

Secondly, our sample does not cover the entire diversity of galaxies, hence there are some missing links that could create a branch of low mass objects emerging after for example C3. The two groups C9 and C10 could then be relics of this distinct evolutionary path where mass does not increase as much as in the main trend seen from our sample. We note that such drops in the evolution in the dynamical mass have been found on smaller samples by Fraix-Burnet et al. (2010, 2012).

Lastly, the low masses of C9 and C10 can be a real loss of mass of these relatively evolved galaxies. The C9 and C10 galaxies are small and faint, but they differ in $B - V$, n_V , ϵ , and T , C9 being made of blue lenticulars and C10 of red ellipticals. C10 galaxies are more central and C9 formed more stars recently (higher $sfr1$ and $sfr2$). Hence, there might be some indication that C9 could be stripped-off galaxies (D’Onofrio et al. 2015), but for C10 this is less clear. So why are C9 and C10 placed nearly at the end of the tree with such low masses (and blue colour for C9)? This is because these two groups follow the main trends in three other parameters (increasing n_V and D_{4000} , decreasing $\langle\mu\rangle_e$), while the two remaining parameters, amongst the seven used to build the tree, H_β and $\log(R_e)$, do not seem to bring too many constraints (no obvious potentially strong trend). As a consequence the algorithm found the shortest path between the groups by optimising the main trends for the maximum of parameters.

Groups C8 and C9 belong to the same ensemble of branches (C6 to C9) and separate at the same node. C8 is made of late-type galaxies, more massive than C9, but as bluish as C9. This seems to be consistent with C8 having “just entered” the cluster environment, possibly not yet as stripped-off as C9. It turns out that C8 is one of the groups that hosts jellyfish galaxies (Sect. 5.7).

Regarding the field sample, the global increase in mass is largely present but there are several departures that have a wiggling shape (Figs. 7 and 10). Contrary to the cluster sample, only one parameter (n_V) amongst the seven used for the analysis shows a regular trend, although it is not perfectly monotonic. This means that the algorithm was not able to find the most parsimonious solution with several regularly increasing or decreasing parameters, so it could not avoid drops in the mass evolution (known as regressions in MP analysis).

Groups C11 and C12 have high masses, are red and ellipticals, and still form some stars (average $sfr1$ and $sfr2$) even if much less than earlier on (higher $sfr3$ and $sfr4$). Clearly they are the most central galaxies. We find such kinds of galaxies in the field sample (F8, F11, F12, F13, and F15) with high masses that are red with an elliptical morphology, but their $sfr3$ is not very much higher than the other groups, indicating that they formed stars only at very early stages ($sfr4$). In addition,

these field sample galaxies have significantly higher sfr^* at all epochs. So these high-mass galaxies, supposedly situated at the end of the evolutionary process of galaxies, have built up very early, and field galaxies appear to form stars at a higher rate than the cluster ones.

5.5. Evolution in size

The sizes of the cluster and field galaxies do not show any evolutionary trend in $\log(R_e)$ (Figs. 6 and 7), but the Sersic index n_V increases and the morphology T decreases. The effective surface brightness $\langle\mu\rangle_e$ becomes brighter in both samples, with a high peak at F11 and F12. So galaxies become more concentrated, more elliptical, and much brighter, whatever their radius and their environment.

The expected evolution in radius with redshift is likely to be not visible for the small range in redshift covered by our samples ($0.02 < z < 0.44$). Hydrodynamical simulations suggest in fact a progressive decrease of R_e going from redshift 4 to 0 (D’Onofrio et al. 2019).

The present data seem then to suggest that the environment has no global effect on the structure of galaxies, where for structure here we mean the effective radius and the Sersic index of galaxies. There is, on the other hand, a hint of morphological segregation in the cluster sample since the early types are at smaller Δ_{CC} . This is the well-known morphology-density relation (Dressler et al. 1987; Fasano et al. 2012). The systematic difference, however, is modest, and the scatter is too large to make a definite inference.

5.6. Emission line properties: Gas deficit

Figures 11–13 show the behaviour of the equivalent width of $H\alpha$ versus the ratio $[NII]\lambda 6584/H\alpha$ for the full, cluster, and field samples, respectively. The diagrams relate the $W(H\alpha)$, a measure of the amount of gas per unit stellar mass, to the ratio $[NII]\lambda 6584/H\alpha$ (Cid Fernandes et al. 2010), which is a tracer of the ionisation mechanism. It cannot be ≥ 0.65 for pure HII regions following the diagnostic diagrams of Kauffmann et al. (2003). Typical HII nuclei have $[NII]\lambda 6584/H\alpha \lesssim 0.4$, LINERs and Seyfert-type active galactic nuclei (AGN) spectra have $[NII]\lambda 6584/H\alpha \gtrsim 0.7$.

There is apparently a progression in the three samples from top left to bottom right of the diagram, confirmed by the evolution of the medians of the cluster and field sample groups (Figs. 14 and 15). This means that there are less HII regions with diversification. For the cluster sample, there are more HII regions of galaxies in the first two groups (C1–2) and in C8 and C9. Groups C1, 2, 8, and 9 mainly make up the blue sequence of the cluster galaxies (see below). Groups C3, 4, 6, and 10 are made predominantly of transition objects between HII and LINERs. C11 and 12 are definitely dominated by LINER spectra and transition objects (e.g. HII galaxies are fully absent). However, there are departures from this trend, most notably C8 and C9. However, C8 is predominantly made of late-type galaxies, and C9 seems to be the lower mass counterpart of C8. Both show high $W(H\alpha)$ and hence large sfr . It is interesting to recall that C6 to C9 and C10 to C12 define two distinct ensembles on the tree (Fig. 5), which we suspect to have increased their stellar mass in two different ways, the former ensemble by forming stars and the latter by accreting old stars (see Sect. 4.2) in agreement with the line emission results.

This trend – that is, passing from HII dominated samples to AGN/LINER spectra through intermediate objects – can also be

Fig. 11. Equivalent width of $H\alpha$ vs. $[NII]/H\alpha$. The vertical dotted lines correspond to $[NII]/H\alpha = 0.4$ and 0.65 and the horizontal one is for $W(H\alpha) = 2.5$ (see text). Dots are for cluster galaxies and crosses for field ones.

Fig. 12. Idem Fig. 11 for cluster galaxies.

seen in the F groups, with F11 and F15 the only groups strongly departing from it. Groups F11 and 15 isolate groups of early-type galaxies with a population of HII emitters.

However, there is a notable difference between field and cluster sample groups: in the latter, the $W(H\alpha)$ amongst field galaxies is higher by a factor of a few than amongst cluster galaxies except for groups C8 and C9 (Figs. 14 and 15). A large fraction of galaxies in the groups not dominated by HII nuclei falls below the limit of $W(H\alpha) = 2.5 \text{ \AA}$ that separates weak emitters from “retired” and quiescent sources according to Cid Fernandes et al. (2010).

A comparison between groups confirms the suggestion of Marziani et al. (2017) about a lower amount of gas in cluster galaxies, and of a population of weak, non-HII line emitters with high $[NII]\lambda 6584/H\alpha$ associated with the red sequence of galaxies.

Fig. 13. Idem Fig. 11 for field galaxies.

Fig. 14. Evolution of the median of $W(H\alpha)$ for the cluster sample. Shaded regions show 1σ dispersion around median. The horizontal dotted line is for $W(H\alpha) = 2.5$ (see text and Figs. 11–13).

5.7. Jellyfish galaxies

A number of “jellyfish galaxies” have been identified in the WINGS database by Poggianti et al. (2016). These objects are outstanding examples of galaxies that show knotted arms of material resembling the tentacles of jellyfish. They are believed to be galaxies that are still gas rich and have just entered the cluster environment.

Jellyfish galaxies are found mainly in C2 and C8, with a prevalence of $\sim 10\%$ in each group, and in F3, F9, and F13. Groups F3 and F13 are large, and so the prevalence of jellyfish galaxies is very low, $\sim 2\%$. In F9 they are $\sim 10\%$ of the sample. Amongst the cluster groups, C2 is dominated by HII regions while C8 is shifted towards intermediate objects. However, both groups are predominantly above the limit of retired

Fig. 15. Idem Fig. 14 for the field sample.

Fig. 16. Histograms for jellyfish galaxies that are in our cluster (black lines) or field (grey lines) samples, for the parameter δ (see text) and the distance Δ_{cc} to cluster centre.

galaxies. The group assignment is consistent with the idea that jellyfish galaxies are characterised by ram pressure stripping of gas by the intra-cluster medium, triggering starbursts along the tails that give jellyfish galaxies their characteristic morphology. It is also consistent with the idea that jellyfish galaxies are just entering the cluster environment.

It is remarkable that some (18) jellyfish galaxies are revealed in the field sample. This is apparently surprising since jellyfish morphology is believed to be exclusive to the cluster environment. Their cluster-centric distance with respect to jellyfish galaxies classified as cluster members is only slightly skewed towards larger distances; the non-member distribution is well within the distribution of cluster-centric distances for jellyfish galaxies that are cluster members. The field sample galaxies do not meet a radial velocity criterion that implies that their radial velocity cz should be different from the $cz_{cluster}$ by less than three times the velocity dispersion of the cluster. Looking at the distribution of the variable $\delta = |cz - cz_{cluster}| / \sigma$, we note that there are about six galaxies that are within $\delta \lesssim 6\sigma$ (Fig. 16). Considering that jellyfish galaxies are probably high velocity intruders (Jaffé et al. 2018), these galaxies could be physically associated with the cluster. The remaining jellyfish have δ variables that are too large for a physical association. Even if their nature remains to be clarified on a source-by-source basis, they may be true jellyfish galaxies occurring in groups or low-mass halos (Poggianti et al. 2016), or be due to gravitational interactions amongst field

Fig. 17. Scatter plots between n_v , $\log(M^*)$, and M/L for cluster (left column) and field (right column) galaxies.

galaxies that are known to lead to tidal tails with star-forming knots (e.g. [Duc 2013](#)), loosely resembling the tails of cluster jellyfish.

5.8. The $\log(M^*)$, n_v , and M/L relation

A relation between the stellar mass M , n_v , and M/L has been found by [D’Onofrio et al. \(2011\)](#). We confirm this correlation (Fig. 17) for the cluster and field samples with a clear sequence along the tree. Not much difference is present between the two samples. However, the correlation does not hold within most of the groups, suggesting that this correlation is probably due to the co-evolution of these variables that is significantly influenced by the evolutionary history of the galaxies.

The shape of galaxies and their stellar population are therefore linked to each other statistically. Nature does not permit the existence of blue ellipticals and red spirals at the present epoch. This relation is therefore another example of co-evolution amongst galactic components, the most famous being that between the black-hole mass and the bulge mass ([Magorrian et al. 1998](#)).

The existence of co-evolving components is particularly interesting in a framework where galaxies form progressively through merging events that are necessarily random in their nature. This implies that behind causality there is a well-defined evolving flux of galaxy structures. The memory of disturbing events, due for instance to minor mergers, should be rapidly lost, while in the case of catastrophic events such as major mergers,

Fig. 18. Effective radius vs. mass for the groups of the full sample. Grey points are the galaxies of the full sample.

galaxies enter a new configuration that rapidly forgets the previous galaxy properties. Co-evolution is probably at the origin of many of the observed scaling relations.

5.9. The $\log(M^*)$ versus $\log(R_e)$ relation

The $\log(R_e)$ versus $\log(M^*)$ scatter plot shows that the galaxies in A15 and C12 are the only ones that define a clear mass-radius relation (Figs. 18 and 19). These are the most diversified groups with the most massive objects, in good agreement with numerical simulations ([Taylor & Kobayashi 2016](#)). However, this is not true for the field sample where no such relation is present in any of the groups (Fig. 20). The group distribution is highly discriminated for the three samples, and is identical to the one found in [Fraix-Burnet et al. \(2010, 2012\)](#). This can be explained by the different assembly histories of the galaxies, through transformation events such as mergers (e.g. [Robertson et al. 2006](#)) or AGN feedback ([Taylor & Kobayashi 2016](#)). We note that the sample studied by [Fraix-Burnet et al. \(2010\)](#) is made of cluster galaxies only (699 early-type galaxies in 56 clusters, [Hudson et al. 2001](#)), and the one studied by [Fraix-Burnet et al. \(2012\)](#) is a priori a mixture of field, group, and cluster galaxies (424 galaxies from [Ogando et al. 2008](#)) whose proportion is unfortunately unknown to us.

The almost complete absence of a linear relation between $\log(R_e)$ and $\log(M^*)$ for field galaxies is very important. It is a clear suggestion that the relation has its origin in the peculiar cluster environment. The large number of dry merging events should be the key physical mechanism producing large radii for big mass objects. Such events should be almost absent in the field.

The $\log(R_e)$ versus $\log(M^*)$ relation appears almost flat for galaxies with $M^* < 10^{10} M_\odot$ both in ETGs and late-type galaxies (LTGs). This is also seen in numerical simulations, such as Illustris ([Vogelsberger et al. 2014](#)). The linear relation appears only for the most massive galaxies in clusters, where the radius likely increases for the high number of minor merging events while the stellar mass remains almost the same. The linear

Fig. 19. Effective radius vs. mass for the groups of the cluster sample. Grey points are the galaxies of the cluster sample only.

Fig. 20. Effective radius vs. mass for the groups of the field sample. Grey points are the galaxies of the field sample only.

relation $\log(R_e) - \log(M^*)$ is expected for the total mass M , which is dark matter plus baryonic mass. Many effects can determine the absence of such a relation. The most likely explanation, according to models, is that winds from supernovae and feedback effects from the central AGN have expelled most of the gas and quenched the star formation, disconnecting R_e , which depends on the dimension of the potential well of the total mass, from the stellar mass.

The F12 group is quite peculiar, being made up of small size, high mass ($\approx 10^{11} M_\odot$) elliptical galaxies, with an anomalous high D_{4000} and the highest $sfr4$. These objects are rare in other surveys at different redshifts, such as that of Fornax (Venhola et al. 2018), that of isolated galaxies (Fernández Lorenzo et al. 2013), the SDSS (Shen et al. 2003; Nair & Abraham 2010), the Galaxy And Mass Assembly survey (GAMA; Lange et al. 2015) and the survey done with the Sydney Australian Multi-object Integral-Field Spectrograph SAMI (Scott et al. 2017). However,

Fig. 21. Effective radius vs. effective surface brightness for the groups of the full sample. Grey points are the galaxies of the full sample.

super-dense galaxies are known to exist either in clusters (Valentinuzzi et al. 2010) or in the field (Poggianti et al. 2013). In the field the frequency of these objects with masses larger than $10^{10} M_\odot$ is estimated to be around 4.4%. In our sample we have 24 members of the F12 groups over a total number of field galaxies of 1158 ($\sim 2\%$ frequency). If we add the F11 group, which is also made of quite massive and small objects, we reach a frequency of $\sim 5\%$. These values are in line with the frequency estimated by Poggianti et al. (2013), but in any case we prefer to warn the reader about a possible discrepancy with the above cited surveys.

5.10. The Kormendy $\langle \mu \rangle_e$ versus $\log(R_e)$ relation

The Kormendy relation $\langle \mu \rangle_e$ versus $\log(R_e)$ (Kormendy 1977) is visible in the three samples (Figs. 21–23) but with a large dispersion and it is not present in most of the groups. Even if the most diversified groups (A15, C12, F15) show a clear relation, some other groups of massive galaxies do not (A8, A11, A13, C11, F12, F13), while some groups of less massive galaxies do (A3, A10, A14, C1, F14). Mass is thus not the only ingredient necessary for this relation to hold.

In addition, there is a lot of dispersion in this plot when the entire sample is considered, and it is very significantly reduced when only individual groups are considered. This is also striking on the $\log(R_e)$ versus $\log(M^*)$ plots (Sect. 5.9). The reason for this dispersion is the mixture of different populations at different stages of evolution, similar to the emission line diagnostic plots (Sect. 5.6), where the different sources of ionisation can only be separated by a multivariate analysis (or by combining several such diagrams; see e.g. de Souza et al. 2017).

We note that our groups are distributed in this plane without much overlap. For instance, considering the Kormendy diagram for the cluster sample (Fig. 22), the regions to the top right are occupied by the three most massive groups (C8, 11, 12), so the dispersion perpendicular to the global relation is mainly due to mass. This is also true for the full and field samples. This means that by dividing the samples into two mass ranges,

Fig. 22. Effective radius vs. effective surface brightness for the groups of the cluster sample. Gray points are the galaxies of the cluster sample only.

Fig. 23. Effective radius vs effective surface brightness for the groups of the field sample. Grey points are the galaxies of the field sample only.

we find almost parallel distributions. The Kormendy relation is clearly linked to the mass-radius relation (see for more details D’Onofrio et al. 2019).

5.11. Colour-magnitude diagram

The colour-magnitude diagram for the cluster galaxy sample (Fig. 24) shows two main regions that are occupied by groups C1, 2, 8, and 9, which correspond to the blue sequence, and a longer and more populated region corresponding to the red sequence. The projection of the tree onto this diagram (Fig. C.1) depicts the evolutionary path: the evolution first goes vertically between the two regions. After C5, there is a bifurcation corresponding to the two ensembles seen in Fig. 5: one along the red sequence (C10 to C12), and the other one (C6 to C9) going downwards with C8 and C9, clearly in what is often called the

Fig. 24. Colour-magnitude diagram for the cluster sample.

Fig. 25. Colour-magnitude diagram for the field sample.

green valley. Hence, we have a somewhat expected evolution from the blue to the red sequence, and subsequently along the red sequence towards brighter galaxies, but the green valley is not a transitional region, rather it is the destination of another branch of the evolutionary path of galaxies.

The field sample (Fig. 25) is remarkably different, the redder and less luminous part being poorly populated, while the blue and more luminous part is quite dense. There seems to be a shift in a large part of the galaxies between the two diagrams. Even if this might partially be due to a Malmquist bias, it appears that the beginning of the evolution from F1 to F2 is very similar to the case of the cluster sample (Fig. C.2). However, the subsequent trajectory passes through the green valley and then to the upper left part of the red sequence.

This diagram is another example of the influence of the environment on the galaxy properties and their evolution. This is also another good proof that the MP approach is able to reveal the characteristics of the evolution of different galaxy populations. In clusters, the red and blue sequences are better defined and we can clearly identify the class of objects in the green valley, which are likely rejuvenated by episodes of star formation or wet merger events. We note in particular the extended sequence of red and faint objects present in the cluster sample. These are almost absent in the field. These objects, once they enter the cluster environment, rapidly lose their gas and stop their star formation, becoming red. In contrast, luminous and blue galaxies are numerous only in the field sample, probably because in clusters star formation has been hindered or suppressed.

6. Conclusions

This study aims to investigate the effect of the cluster environment on the evolution of galaxies. We have used a sample of homogeneous data from the WINGS catalogue, with about 2600 galaxies that are members of clusters and about 1500 ones that are field objects.

We used maximum parsimony, a phylogenetic technique also known as cladistics. It is designed to find evolutionary paths between objects or classes and has now been often used in astrophysics. However, the sample is too large for this technique to be applied directly. We reduced the amount of objects by performing a pre-clustering using a hierarchical clustering method in order to reduce the data set according to its extent and not to its density distribution. We have found that 300 pre-clusters is a good compromise between being representative and the computational constraints of the MP analysis. We used seven parameters for these two machine learning computations: $B-V$, $\log(R_e)$, n_V , $\langle\mu\rangle_e$, H_β , D_{4000} , and $\log(M^*)$. They are all fully documented, and we have checked that there is no redundancy or strong correlation in this set.

We have obtained an evolutionary tree for the two samples both combined and separately. The result with all galaxies does not show distinct evolutionary paths for cluster and field galaxies. Rather, it shows that the cluster environment may somehow accelerate the evolution of galaxies by making them more diversified (different) from primitive galaxies. Also, field galaxies appear more massive at similar evolutionary stages and form more stars. On the tree for the cluster sample, there is a distinct sub-structure, a distinct evolutionary path, which seems to gather rejuvenated or stripped-off galaxies, and goes into the green valley (or even blue sequence) from the red sequence on the colour-magnitude diagram.

The two separate analyses reveal subtle evolutionary differences. Cluster galaxies appear more homogeneous since the groups of the field sample are characterised by slightly more specific and varied properties. The classifications we obtained from the cluster and field trees are only partly driven by mass and colour. The Sersic index is highly correlated with the groups for the field galaxies, and it appears to be one of the main drivers of the evolutionary tree structure together with the effective radius and the stellar content, the mass being here not important at all.

The classification obtained from the tree of the cluster sample does not appear to depend on the clusters, implying some universality to the diversity of cluster galaxies. However, our sample is at low redshift, and the number of galaxies in each cluster is sometimes limited.

An important outcome of a multivariate analysis is the investigation of scaling relations, and in particular the understanding of their dispersion. We find that many, if not all, such relations are explained by co-evolution, and that a proper classification makes these relations appear as a mixture of galaxy populations at different stages of evolution. For instance, we show that the dispersion in the Kormendy relation is in fact due mainly to at least two parallel relations with two different masses. Also the Kormendy relation is not due to mass since it appears in some low mass groups as well. Also, the mass versus radius relation holds only for highly diversified and massive groups, but not for the field sample.

Finally, we find a striking difference in the colour-magnitude diagrams for cluster and field galaxies, and the evolutionary trees projected on this diagram shows that evolution does not go simply from the blue sequence to the red one. Even if it is true for some cluster galaxies, the others tend to go back to the blue sequence, through the green valley. This latter process is the one chosen by field galaxies.

This paper is another demonstration that unsupervised machine learning clustering is able to identify known classes of objects by making the properties precise, and to highlight more peculiar types of objects. Importantly, the phylogenetic approach is able to provide evolutionary scenarios to explain the physical origins of all these classes. Of course, investigating the diversification of galaxies at a redshift of nearly zero is limited and frustrating. However, we have shown that the evolutionary paths are somewhat complicated: even though galaxies globally become redder, more massive, and less disc-like, there are many other possible trajectories, such as field galaxies, which are bluer and more concentrated but more massive than cluster ones, or cluster galaxies that are stripped-off. Also, there are many more properties of galaxies other than colour, mass, and morphology. What we need now is to go to higher redshifts, trying to find more primitive galaxies and better map the evolutionary paths of present day objects. It is important to recall that even if cluster galaxies undergo more frequent interactions than field ones, the latter certainly were subject to such transforming events a long time ago.

Acknowledgements. We thank the referee for thorough and constructive comments that improved significantly this paper.

References

- Abazajian, K., Adelman-McCarthy, J. K., Agüeros, M. A., et al. 2003, *AJ*, **126**, 2081
- Bacon, R., Copin, Y., Monnet, G., et al. 2001, *MNRAS*, **326**, 23
- Barrow, J. D., Bhavsar, S. P., & Sonoda, D. H. 1985, *MNRAS*, **216**, 17
- Bertin, E., & Arnouts, S. 1996, *A&AS*, **117**, 393
- Bhavsar, S. P., & Splinter, R. J. 1996, *MNRAS*, **282**, 1461
- Blanco-Cuaresma, S., & Fraix-Burnet, D. 2018, *A&A*, **618**, A65
- Cappellari, M., McDermid, R. M., Alatalo, K., et al. 2012, *Nature*, **484**, 485
- Cava, A., Bettoni, D., Poggianti, B. M., et al. 2009, *A&A*, **495**, 707
- Chang, W.-C. 1983, *Appl. Stat.*, **32**, 267
- Cid Fernandes, R., Stasińska, G., Schlickmann, M. S., et al. 2010, *MNRAS*, **403**, 1036
- de Souza, R. S., Dantas, M. L. L., Costa-Duarte, M. V., et al. 2017, *MNRAS*, **472**, 2808
- Dekel, A., & Birnboim, Y. 2006, *MNRAS*, **368**, 2
- Djorgovski, S., & Davis, M. 1987, *ApJ*, **313**, 59
- D'Onofrio, M., Valentinuzzi, T., Fasano, G., et al. 2011, *ApJ*, **727**, L6
- D'Onofrio, M., Bindoni, D., Fasano, G., et al. 2014, *A&A*, **572**, A87
- D'Onofrio, M., Marziani, P., & Buson, L. 2015, *Front. Astron. Space Sci.*, **2**, 4
- D'Onofrio, M., Chiosi, C., Sciaratta, M., & Marziani, P. 2019, ArXiv e-prints [arXiv:1907.09367]
- Dressler, A., Lynden-Bell, D., Burstein, D., et al. 1987, *ApJ*, **313**, 42

- Duc, P. A., & Renaud, F. 2013, in [Lecture Notes in Physics](#), eds. J. Souchay, S. Mathis, & T. Tokieda (Berlin: Springer Verlag), 861, 327
- Dutton, A. A., van den Bosch, F. C., Faber, S. M., et al. 2011, [MNRAS](#), **410**, 1660
- Faber, S. M. 1973, [ApJ](#), **179**, 731
- Faber, S. M., & Jackson, R. E. 1976, [ApJ](#), **204**, 668
- Fall, S. M., & Romanowsky, A. J. 2013, [ApJ](#), **769**, L26
- Farris, J. S. 1989, [Cladistics](#), **5**, 417
- Fasano, G., Marmo, C., Varela, J., et al. 2006, [A&A](#), **445**, 805
- Fasano, G., Vanzella, E., Dressler, A., et al. 2012, [MNRAS](#), **420**, 926
- Felsenstein, J. 1984, in [Cladistics: Perspectives on the Reconstruction of Evolutionary History](#), eds. T. Duncan, & T. Stuessy (New York: Columbia University Press), 169
- Fernández Lorenzo, M., Sulentic, J., Verdes-Montenegro, L., & Argudo-Fernández, M. 2013, [MNRAS](#), **434**, 325
- Fraix-Burnet, D. 2011, [MNRAS](#), **416**, L36
- Fraix-Burnet, D. 2016, in [Statistics for Astrophysics: Clustering and Classification](#), eds. D. Fraix-Burnet, & S. Girard (EDP Sciences), [EAS Publ. Ser.](#), **77**, 221
- Fraix-Burnet, D., Choler, P., Douzery, E., & Verhamme, A. 2006a, [J. Classif.](#), **23**, 31
- Fraix-Burnet, D., Douzery, E., Choler, P., & Verhamme, A. 2006b, [J. Classif.](#), **23**, 57
- Fraix-Burnet, D., Dugué, M., Chattopadhyay, T., Chattopadhyay, A. K., & Davoust, E. 2010, [MNRAS](#), **407**, 2207
- Fraix-Burnet, D., Chattopadhyay, T., Chattopadhyay, A. K., Davoust, E., & Thuillard, M. 2012, [A&A](#), **545**, A80
- Fraix-Burnet, D., Thuillard, M., & Chattopadhyay, A. K. 2015, [Front. Astron. Space Sci.](#), **2**
- Fritz, J., Poggianti, B. M., Bettoni, D., et al. 2007, [A&A](#), **470**, 137
- Fritz, J., Poggianti, B. M., Cava, A., et al. 2011, [A&A](#), **526**, A45
- Gascuel, O., & Steel, M. 2006, [Mol. Biol. Evol.](#), **23**, 1997
- Genel, S., Vogelsberger, M., Springel, V., et al. 2014, [MNRAS](#), **445**, 175
- Hendy, M., & Penny, D. 1982, [Math. Biosci.](#), **59**, 277
- Hennig, W. 1965, [Annu. Rev. Entomol.](#), **10**, 97
- Holt, T. R., Brown, A. J., Nesvorný, D., Horner, J., & Carter, B. 2018, [ApJ](#), **859**, 97
- Hudson, M. J., Lucey, J. R., Smith, R. J., Schlegel, D. J., & Davies, R. L. 2001, [MNRAS](#), **327**, 265
- Jaffé, Y. L., Poggianti, B. M., Moretti, A., et al. 2018, [MNRAS](#), **476**, 4753
- Jofre, P., Das, P., Bertranpetit, J., & Foley, R. 2017, [MNRAS](#), **467**, 1140
- Kauffmann, G., Heckman, T. M., Tremonti, C., et al. 2003, [MNRAS](#), **346**, 1055
- Kormendy, J. 1977, [ApJ](#), **218**, 333
- Lagos, C. D. P., Theuns, T., Schaye, J., et al. 2016, [MNRAS](#), **459**, 2632
- Lange, R., Driver, S. P., Robotham, A. S. G., et al. 2015, [MNRAS](#), **447**, 2603
- Maddison, W. 1991, [Syst. Zool.](#), **40**, 304
- Magorrian, J., Tremaine, S., Richstone, D., et al. 1998, [AJ](#), **115**, 2285
- Marziani, P., D'Onofrio, M., Bettoni, D., et al. 2017, [A&A](#), **599**, A83
- Moretti, A., Poggianti, B. M., Fasano, G., et al. 2014, [A&A](#), **564**, A138
- Murtagh, F., & Legendre, P. 2014, [J. Classif.](#), **31**, 274
- Nair, P. B., & Abraham, R. G. 2010, [ApJS](#), **186**, 427
- Ogando, R. L. C., Maia, M. A. G., Pellegrini, P. S., & da Costa, L. N. 2008, [AJ](#), **135**, 2424
- Poggianti, B. M., Calvi, R., Bindoni, D., et al. 2013, [ApJ](#), **762**, 77
- Poggianti, B. M., Fasano, G., Omizzolo, A., et al. 2016, [AJ](#), **151**, 78
- Ramella, M., Biviano, A., Pisani, A., et al. 2007, [A&A](#), **470**, 39
- Robertson, B., Cox, T. J., Hernquist, L., et al. 2006, [ApJ](#), **641**, 21
- Saitou, N., & Nei, M. 1987, [Mol. Biol. Evol.](#), **4**, 406
- Scott, N., Brough, S., Croom, S. M., et al. 2017, [MNRAS](#), **472**, 2833
- Shen, S., Mo, H. J., White, S. D. M., et al. 2003, [MNRAS](#), **343**, 978
- Snyder, G. F., Torrey, P., Lotz, J. M., et al. 2015, [MNRAS](#), **454**, 1886
- Swofford, D. L. 2003, [PAUP*: Phylogenetic Analysis Using Parsimony \(*and Other Methods\)](#) (Sunderland, Massachusetts: Sinauer Associates)
- Taylor, P., & Kobayashi, C. 2016, [MNRAS](#), **463**, 2465
- Tully, R. B., & Fisher, J. R. 1977, [A&A](#), **54**, 661
- Valentinuzzi, T., Fritz, J., Poggianti, B. M., et al. 2010, [ApJ](#), **712**, 226
- Valentinuzzi, T., Poggianti, B. M., Fasano, G., et al. 2011, [A&A](#), **536**, A34
- Varela, J., D'Onofrio, M., Marmo, C., et al. 2009, [A&A](#), **497**, 667
- Venhola, A., Peletier, R., Laurikainen, E., et al. 2018, [A&A](#), **620**, A165
- Vogelsberger, M., Genel, S., Springel, V., et al. 2014, [MNRAS](#), **444**, 1518

Appendix A: Number of hierarchical clusters and tree rooting

In this appendix, we investigate the influence of the choice of the number of pre-clusters and the tree rooting on the result of the MP analysis.

We have explained in Sects. 3.2 and 3.3 how the number of pre-clusters and the contours of the phylogenetic groups were chosen. We want to show here the robustness of our results with respect to these choices. For this purpose we have performed the same cladistic analysis for the full sample with 100 and 200 pre-clusters.

Due the hierarchical nature of their definition, the three sets of pre-clusters match: a pre-cluster of the 100 set contains one to several pre-clusters of the 200 set and so forth (see Fig. A.1). As a consequence, the medians of the parameters used for the MP analysis similarly span the entire variance of the full sample in the three cases. We can thus expect that the evolutionary paths (the relationships between classes) should not be very different.

This is confirmed on the MP trees obtained using the three pre-cluster sets (Fig. A.2): for each tree, and for each pre-cluster (at tree tips), we show the distribution of galaxies in the groups defined in this paper from the tree in Fig. 3, reproduced here on the right, using the same colour code. The tree shapes in the three cases are not identical, but most of the groups can be well identified on the trees of the 100 (left) and 200 (right) pre-cluster sets, indicating that the majority of the galaxies belonging to one of the A groups remain together. The agreement is better for the 200 pre-cluster tree, although part of groups A5 and A10 are displaced.

The colours on the three trees clearly show that the diversification scheme is very similar as well. By defining groups from the 100 and 200 pre-cluster trees, the violin plots (Figs. A.3 and A.4) confirm this similarity when compared with Fig. 4. This implies that the main conclusions of our study are not modified.

The MP analysis yields an unrooted tree since no direction of changes of the parameters has been provided. To interpret the tree as an evolutionary scenario, an arrow of diversification must be defined by choosing a root for the tree. This root must be as similar as possible as the common ancestor to the objects under study. In this paper, we have used three parameters the evolution of which is thought to be well understood: colour, driven by the age of the stellar population as well and metallicity, Sersic index, which represent the dynamics, and the mass, which tends to increase. The choice of the root can always been disputed, and only the consistency of the final tree can in principle support it or not.

Fig. A.1. Dendrogram of the hierarchical clustering algorithm for the full sample. The horizontal red lines indicate from top to bottom the cuts defining the 100, 200, and 300 pre-clusters, respectively.

To provide astronomers an idea of the consequences of a different root, we show here a tree rooted with group A7 (Fig. A.5) with the corresponding sequence of violin plots (Fig. A.6). The tree shows two main branches splitting at group A7, each representing a kind of lineage. Interestingly, the proportion of cluster galaxies is on average significantly higher in the branch A8 to A15 than in the branch A6 to A1 (Table 2) although this is not a clear-cut separation.

The sequence of the changes of the variables along the tree (Fig. A.6) should be read with caution, the two branches must be seen as two parallel separate evolutionary paths. The striking behaviour is the decreasing colour, that is, galaxies become bluer, especially in branch A6–A2, which is not compatible with the ageing of stars and the increasing metallicity. This fact by itself rules out the choice of group A7 as root. The mass is higher in branch A8 to A15 than in A7, but is slightly lower in branch A6 to A2. The morphology also in the second branch clearly evolves towards spiral types. We think that the changes within each branch are not in good agreement with our knowledge of the physics of the galaxy evolution. One could have chosen for instance only mass to root the tree, using group A3, but this would not have significantly changed our result, and it is doubtful that the evolution (and the diversification) of galaxies would depend on only one single property.

Fig. A.2. Trees obtained for the full sample with 100 (*left*), 200 (*middle*), and 300 (*right*) pre-clusters (see Sect. 3.2). The tree on the right is the one shown in Fig. 3. The coloured bars at each tip of the trees give the distribution of the galaxies of the corresponding pre-cluster in the groups defined in this paper (the classes A).

Fig. A.3. Violin plots for groups defined on the tree with 100 pre-clusters (left in Fig. A.2) as described in Sect. 3.3. This figure should be compared with Figs. 4 and A.4. The violin plot of group 7 for parameter D_{4000} spreads between 5 and 9 and is thus well above the diagram.

Fig. A.4. Violin plots for groups defined on the tree with 200 pre-clusters (middle in Fig. A.2) as described in Sect. 3.3. This figure should be compared with Figs. 4 and A.3.

Fig. A.5. Same tree as in Fig. 3 for the full sample but rooted with group A7. The colours and names of the groups are identical.

Fig. A.6. Same plot as in Fig. 4 for the full sample but rooted with group A7 according to tree in Fig. A.5. The colours and names of the groups are identical.

Appendix B: Statistical correlations amongst the variables

These figures are described in Sect. 5.1.

Fig. B.1. Principal component analysis using the seven parameters used for the clustering and MP analyses (Sect. 5.1). *Top left:* scree plot for the full sample showing the variance (eigenvalues) of the seven principal components (PCs). The curve represents the cumulative percentage of the total variance. The scree plots for the cluster and field sample are very similar. *Top right:* projection of the data on the plane of the first two PCs with arrows showing the loadings and direction of the parameters. The same diagram is plotted for the cluster (*bottom left*) and field (*bottom right*) samples.

Fig. B.2. Correlation (absolute value) of the seven variables with the groups (solid lines) and the pre-clusters (dashed lines; Sect. 5.1).

Appendix C: Projection of the tree on the colour-magnitude diagram

These figures are described in Sect. 5.11. The trees in Figs. 24 and 25 are projected on the bivariate plot $B - V$ versus M_V .

The tips of the trees are the medians of these parameters for each group. The internal nodes are given values by minimising the squared changes of the parameters along the branches (squared-change parsimony ancestral state reconstruction, Maddison 1991).

Fig. C.1. Tree projected onto the CMD diagram for the cluster sample.

Fig. C.2. Tree projected onto the CMD diagram for the field sample. The median values of the groups F3 and F6 are superimposed on this plot.