A Maximum Parsimony analysis of the effect of the environment on the evolution of galaxies

Didier Fraix-Burnet, Mauro d'Onofrio, Paola Marziani

Astronomy & Astrophysics manuscript no. WINGS_MP
July 29, 2019

A Maximum Parsimony analysis of the effect of the environment on the evolution of galaxies

D. Fraix-Burnet[1], M. D'Onofrio[2], and Paola Marziani[3]

[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 April 2019; accepted July 2019

ABSTRACT

Context. Galaxy evolution and the effect of environment are most often studied using scaling relations or some regression analyses around some given property. These approaches however do not take into account the complexity of the physics of the galaxies and their diversification.
Aims. We here investigate the effect of cluster environment on the evolution of galaxies through multivariate unsupervised classification and phylogenetic analyses applied to two relatively large samples from the WINGS survey, 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}, $\log(M^*)$.
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 they are statistically more diversified from a primitive common ancestor. The separate analyses show a hint for a slightly more regular evolution of the variables for the cluster galaxies, which may indicate they are more homogeneous as 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 which gathers rejunevated 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 the simple scaling relations by taking into account several properties together. The phylogenetic approach is invaluable to trace the evolutionary scenarii and project them onto any biavariate diagram without any a priori modelling. Our WINGS galaxies are all at low redshift, and we now need to go to higher redshfits to find more primitve 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 enter 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 the galaxy formation models have still 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 their identification is most often made from trials-and-errors, more rarely from statistical tools such as the principal component analysis (PCA) or the multivariate analyses that are explicitly aimed at minimising 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, showing the complexity of galaxy physics and evolution. Another reason can be that the scaling relations reveal some hidden (latent) variable which 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 indicate the amount of inflow and outflow processes across cosmic epochs, 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, e.g. 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. These information are extremely important for tuning models and simulations that today reproduce quite well the dark matter properties, but fail in reproducing the whole range of observable properties of stellar structures, probably for the complexity of baryon physics at local scales.

In this work we want to explore an alternative approach that we hope can permit to achieve a 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 the phylogenetic approaches, do reveal the equivalent of "scaling relations", but in a space whose dimension is given by the number of variables. This is much harder to visualise with plots representing real data and this is why there exists many visualisation tools (the typical cladistic tree being one of them) but they are still difficult to confront 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 will 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 biases. Good and abundant measures are available today only for nearby galaxies, being the high redshift objects still sparsely observed and at the limit of present day telescopes. However, in the last years thanks to the development of galaxy surveys, such as WINGS (Fasano et al. 2006), SDSS (Abazajian et al. 2003), SAURON (Bacon et al. 2001) and many others, it was possible to collect a lot of galaxy measures 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 the multivariate statistical analysis for checking the possibility of understanding the preferential paths of galaxy evolution hidden within the complex interrelations among 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 $\approx$ 250 Mpc. Future works with data for high redshift galaxies might in a near future complete the construction of the main paths of galaxy evolution.

The paper is designed as follow. In Sec. 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. Sec. 3 introduce the multivariate analysis approach, in particular the Maximum Parsimony method (Sec. 3.1). We discuss the selection of the variables used in the phylogenetic analysis in Sec. 3.4 and we present our results in Sec. 4 discussing three particular cases: the full sample (Sec. 4.1), the cluster sample (Sec. 4.2) and the field galaxy sample (Sec. 4.3). The discussion follows in Sec. 5 and our conclusions in Sec. 6.

2. The data

2.1. The WINGS sample

The data used in this work are those belonging to the *Wide-field Imaging of Nearby Galaxy-clusters Survey* (hereafter WINGS) survey.[1] 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), that 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 north and 22 in the south hemisphere) with the spectrographs WYFFOS@WHT (λrange = 3800 ÷ 7000 Å, resolution FWHM = 3 Å) and 2dF@AAT (λrange = 3600 ÷ 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) have been analysed with spectrophotometric techniques deriving the star formation rate $sfr*$ at different epochs, the stellar masses M^* and ages, the internal extinction A_V and the equivalent widths of the main absorption features (Fritz et al. 2011).

A full re-analysis of the WINGS spectra, aimed at measuring the emission line properties, has been also performed by Marziani et al. (2017). These authors considered 46 clusters of the WINGS-SPE survey, and performed population synthesis from spectra that were not considered in (Fritz et al. 2011): spectrophotometric 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. The adopted sample

From the WINGS database, a sample of 4100 galaxies has been extracted for this work. Distances were determined from their redshifts. 2624 galaxies were determined 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 (see 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 have been analysed by Ramella et al. (2007). Cluster background galaxies have been eliminated on the basis of the colour-magnitude diagrams (see Valentinuzzi et al. 2011) and on

[1] See the WINGS web-site for all the details of this project at http://web.oapd.inaf.it/wings.

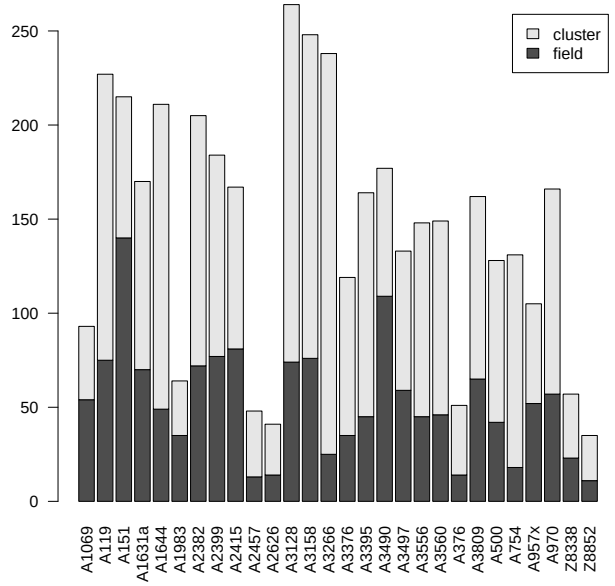

Fig. 1. Histograms of cluster and field sample galaxies into the clusters.

Table 1. Variables available for the WINGS sample. Only the seventh first parameters (top panel) have been used for the clustering and MP analysis.

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 Hβ
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 $M_\odot yrs^{-1}$: stellar mass produced within the age bin $0 - 2 \cdot 10^7$ yrs
$sfr2$	idem of $sfr1$ for the bin $2 \cdot 10^7 - 6 \cdot 10^8$ yrs
$sfr3$	idem of $sfr1$ for the bin $6 \cdot 10^8 - 5.6 \cdot 10^9$ yrs
$sfr4$	idem of $sfr1$ for the bin $5.6 \cdot 10^9 - 17.8 \cdot 10^9$ yrs
$M1/M4$	Stellar production ratio: $\left(sfr1 \times 2 \cdot 10^7\right) / \left(sfr4 \times \left(17.8 \cdot 10^9 - 5.6 \cdot 10^9\right)\right)$
W(Hα)	Rest-frame W(Hα)
[NII] /Hα	Intensity ratio [NII]λ6584/Hα

the results obtained by Ramella et al. (2007). In general, galaxies redder than B–V=1.4, not belonging to background clusters have been kept. The reader should keep in mind this fact that could partly affect our conclusions, when the cluster and field samples are compared.

The GASPHOT tool used in WINGS to derive radius, surface brightness, mass, ellipticity etc, takes into account the convolution with the PSF. Hence, the model profiles includes the PSF measured on nearby stars. In addition the smallest galaxies cover always a surface of 200 pixels squares, so the minimum radius is about 8 pixels while the typical stellar PSF of the survey of

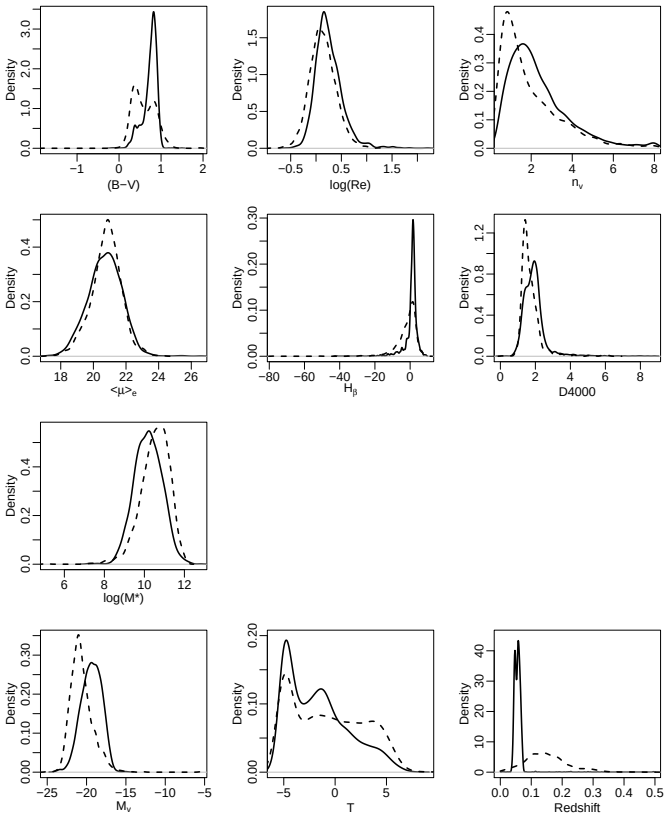

Fig. 2. Gaussian kernel density distribution of the seven variables used in the clustering and MP computation (Table 1), plus the absolute V magnitude, the morphology T and the redshift, for the cluster (solid line) and field (dotted line) samples. The scale on the y-axis is given by the integral of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

around 3 pixels. The light profiles are well 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 most 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 on the top part of Tab. 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 kpc 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 Å D_{4000}; the logarithm of the stellar galaxy mass log(M^*) derived from the SED fitting (Fritz et al. 2007). The distribution 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 have a smaller Sersic index (Fig. 2). This implies that any difference in clustering results between the two populations could not be explained simply some monovariate differences.

The other variables are used for describing the characteristics of the groups and their mutual relationships. Among them, the total galaxy luminosity M_V, the morphological types derived with Morphot 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 BCG, the star formation rates ($sfr1$, $sfr2$, $sfr3$ and $sfr4$) computed in four age ranges ($0 - 2 \cdot 10^7, 2 \cdot 10^7 - 6 \cdot 10^8, 6 \cdot 10^8 - 5.6 \cdot 10^9$ and $5.6 \cdot 10^9 - 17.8 \cdot 10^9$ years), 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 $M1/M4 = \left(sfr1 \times 2 \cdot 10^7\right) / \left(sfr4 \times \left(17.8 \cdot 10^9 - 5.6 \cdot 10^9\right)\right)$), the equivalent width W(Hα) and the intensity ratio [NII]λ6584/Hα[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

3.1. The 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 techniques 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 in 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 heavily used to determine the galaxy clusters in order to map the spatial distribution of the 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.

Among 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 minimizing 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. To each internal node is associated a real value $f(u)$. The score s of a tree equals the sum over all edges of the absolute difference between those values:

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

This can be directly extended to continuous characters or values.

As a consequence, two objects which 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; Fraix-Burnet et al. 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 that we make a consensus tree which keeps internal nodes appearing in at least 70% of them. We aim at what is called a "resolved" tree (or binary tree) where each internal node has three branches and 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 a 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) which thoroughly explores all the possible tree topologies and select 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, say, 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 a non-sense, but a phylogeny, in other word 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

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

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 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 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 much 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 subgraphs. 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 in species, genus, family, etc. The choices depend on the granularity needed for the study and is 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 have 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 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 a complementary information with the relationships between these classes. As a matter of fact, this is the first time that a hierarchical (clustering and phylogenetic) classification is 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 since they 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 in ancestorship and can destroy the tree structure. Since it is not always easy to detect these "flaws" a priori, we are bound to try different subsets and keep the results that show the most parsimonious trees that are as much 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 with $\log(R_e)$ 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 very uncertain because of the orientation effects.

We end up with the seven parameters described on the top part of Table 1: B–V, $\log(R_e)$, n_V, $\langle\mu\rangle_e$, H_β, D_{4000}, $\log(M^*)$. The disregarded variables are also reported on 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 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 Maximum Parsimony 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

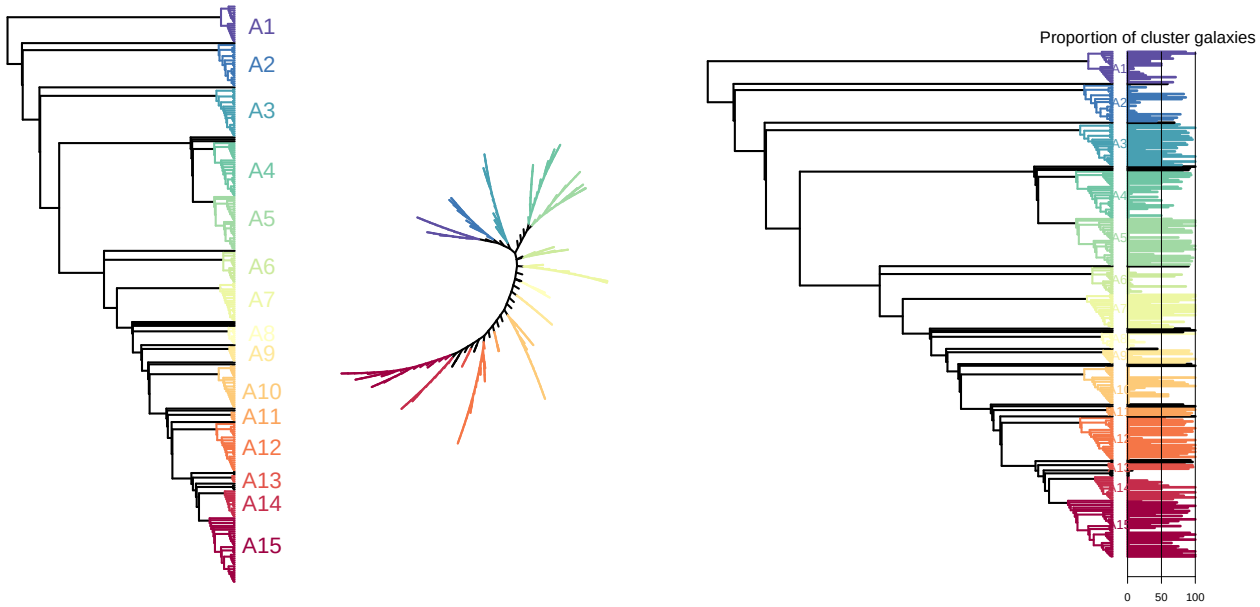

Fig. 3. *(Left)* Tree for the full sample with group definition. Only groups with more than 50 galaxies are considered for this tree in this paper. The tree has been rooted with the group (at the top) having low values of B–V, n_V and log(M*). *(Middle)* Same tree but unrooted. *(Right)* Same tree with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

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 have performed three identical analyses (same seven parameters (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 evolutions diverged, or did they evolve similarly but at a different pace due to their environment? Since we find 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 on 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) having 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 a help for an 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 on this choice. Rooting a tree consists in choosing the group that is considered as the most ancestral and draw the tree starting from this group. We have chosen blue colour, low Sersic index and low mass since their global evolution is known. 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 formed stars very early very low $sfr1$, $sfr2$ 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 which 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 brunches) could have been defined but this would complicate the discussion in the paper. We recall that each tip (leave) of the tree is a pre-cluster (see Sect. 3.2) of one or several galaxies. For reason 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 resp.). 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, 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 of the percent-

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

	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

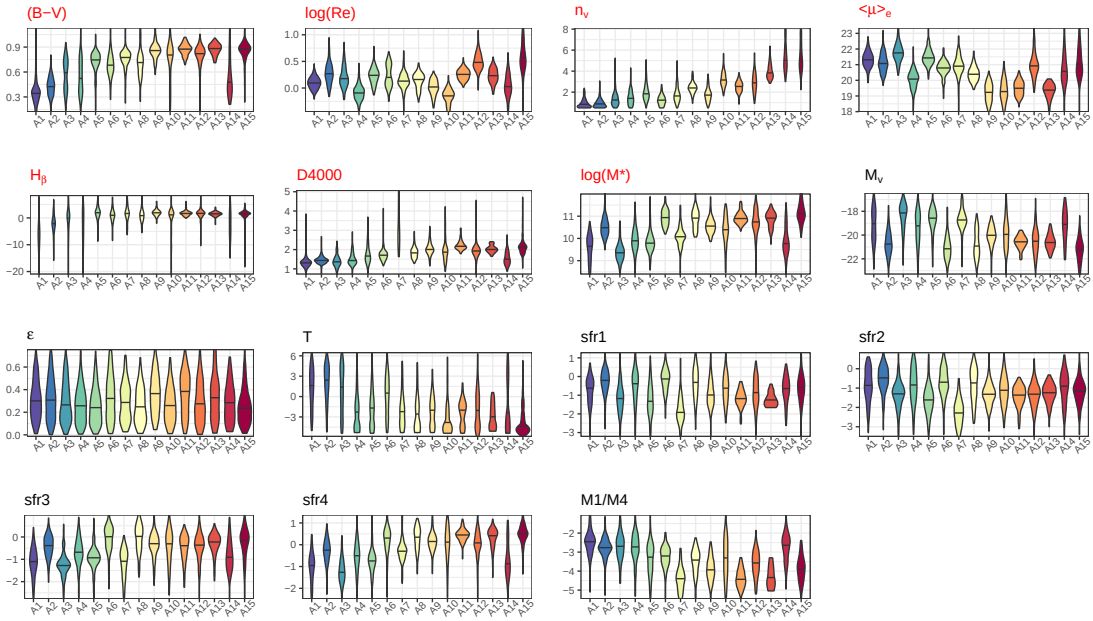

Fig. 4. Violinplots for fifteen variables and the groups A[1-15] of the full sample as defined on Fig. 3. The seven first parameters only (in red) were used for the analysis. The horizontal bar is the median value, and each "violin" symmetrically depicts the distribution of the variable.

age 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.

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 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 that make them suffer from more frequent and/or violent events.

The properties of the groups are represented as violinplots 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. log(R_e) wiggles with no clear evolutionary trend. The Sersic index n_V, also used to root the tree, increases regularly. $\langle\mu\rangle_e$ decreases (the effective surface brightness increases) 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 an opposite manner. ϵ does not show much significant difference between the groups since the intragroup variance is large. The 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 diverse in groups from 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 few exceptions: A6 and A14. Finally, the stellar production ratio

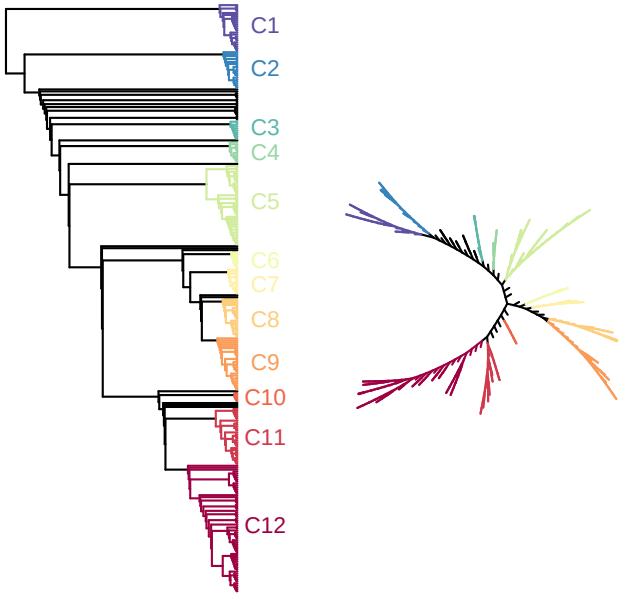

Fig. 5. Tree from the analysis with the cluster sample. Only groups with more than 50 galaxies are identified. The tree on the left has been rooted with the group (at the top) having low values of B–V, n_V and log(M*). To the right, same tree but unrooted.

M1/M4 is lower from A7 to A15, except for A14 and possibly A10.

The properties as given by the violinplots explain the structure of the tree and allows a physical understanding of the groups. The gross picture of galaxy evolution is correctly recovered: galaxies become redder, with ageing stars, more massive, the most diversified galaxies formed stars earlier in the past, they tend to get early-type morphologies. 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 astonishing low B–V, low mass, low D_{4000}, and 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 which have more that 50 galaxies (C1-C12, Fig. 5), leaving 336 (12.8%) galaxies aside.

Three groups dominates (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 not to 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. 6a) 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. Recall that the latter three form a specific ensemble on the tree. Departing from this scheme, C8 is composed by spirals.

It is tempting to associate C6 to A7 (being composed at 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. 6a, 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 are clearly distinct along these two branches. B–V decreases in the first branch from C6 to C9 and increases in the second one from C10 to C12. 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 history evolutions are clearly visible in the stellar production ratio M1/M4 that increases in the first branch and decrease in the second one. Finally, the second branch contain 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. 7) 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 thus leaving 318 (22.5%) galaxies aside (Fig. 7). We note that the group F6 with 31 galaxies is one of the 300 pre-clusters and thus occupy 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. 6b) 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. 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 Fig. 4 and 6a 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 may be 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 in 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 to find a smooth evo-

Table 3. Distribution of galaxies in the different groups from the analysis with the cluster sample. The table also lists the numbers of galaxies in the 27 small groups not considered in this paper.

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	*Total*
100	121	70	124	367	59	160	144	110	51	442	540	2288
Composition of the remaining 27 small groups: 1 1 3 4 7 7 8 8 9 9 9 9 10 10 11 11 12 14 15 16 16 18 19 21 23 24 41												

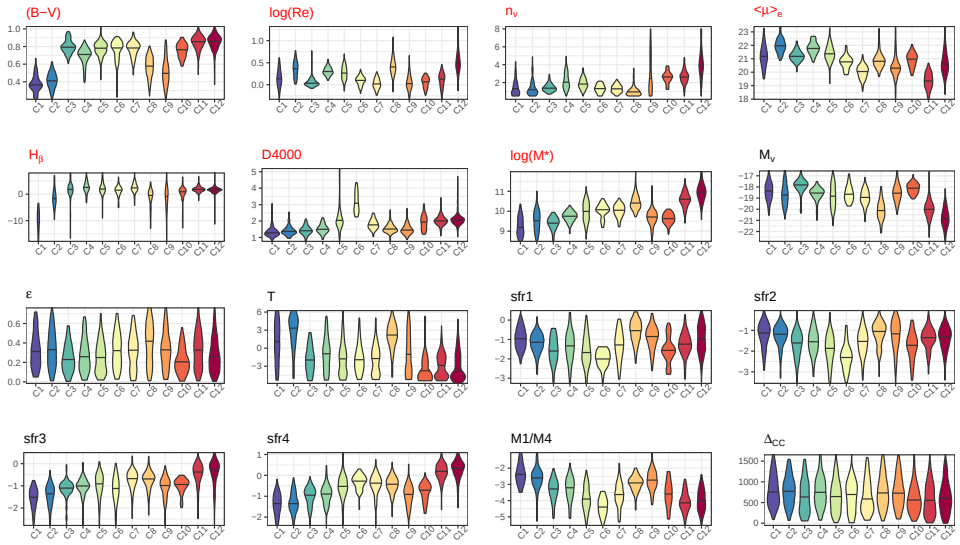

Fig. 6a. Violinplots for sixteen variables and the groups C[1-12] for the cluster sample. Δ_{CC} is the projected distance of a galaxy to the centre of a cluster.

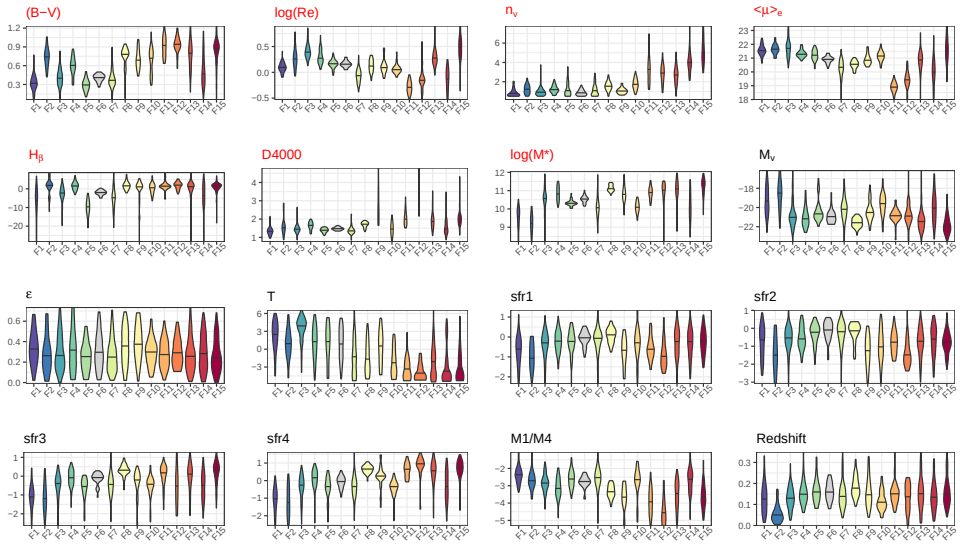

Fig. 6b. Violinplots for sixteen variables and the groups F[1-15] for the field sample.

lutionary 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 magnitudes 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 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 Ser-

Table 4. Distribution of galaxies in the different groups from the analysis with the field sample. The table also lists the numbers of galaxies in the 51 small groups not considered in this paper.

F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	*Total*
75	34	176	53	35	31	219	40	33	40	36	24	182	79	101	1158
Composition of the remaining 50 small groups:															
1 1 1 2 3 3 3 3 3 3 4 4 4 5 5 5 5 5 5 5 5 5 6 6 6 6 7 7 7 7 7 7 7 8 8 8 8 8 9 9 9 9 10 10 10 10 11 11 11 16															

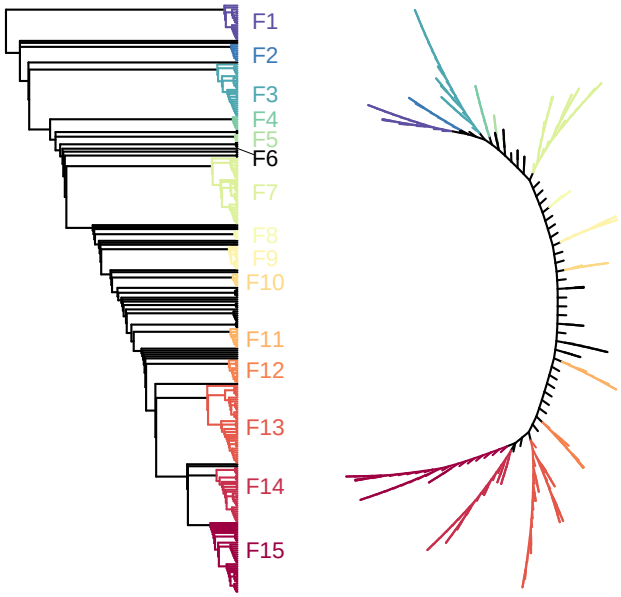

Fig. 7. Tree for the field sample. Only groups with more that 23 galaxies are identified. The tree on the left has been rooted with the group (at the top) having low values of B–V, n_V and log(M*). Group F6 is a pre-cluster and occupies a single branch of the tree indicated by the tick mark. To the right, same tree but unrooted.

sic index and more massive galaxies (Fig. 2) while we assumed that ancestral galaxies were similar but less massive. This suggest that field galaxies evolved more in mass than the 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 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 such age. These types of galaxies are present with higher prevalence among 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. But 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 MP analyses. The Pearson correlation coefficient is always lower than 0.5 for the full sample, 0.57 (log(M*) vs log(R_e)) for the cluster sample and 0.67 ($\langle\mu\rangle_e$ vs log(R_e)) for the field sample. We conclude that there is no significant correlation that could drive 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 Principal Component Analysis (PCA). As shown in Fig. B.1 (top left diagram), there is no really dominant eigenvector and the first four 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 in-between the latter and the five remaining parameters. The conclusion of the PCA analysis is that there is no parameter nor 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 much 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 cor-

relations 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 shows always 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 this 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 6a and 9) 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 of the same order as group C1. In addition, since the groups have been identify in the seven-dimension 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) synthetises 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, showing 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 probably not independent. The classification is not influenced by some particular properties, but there is a clear difference between the cluster and field samples about 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, 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.

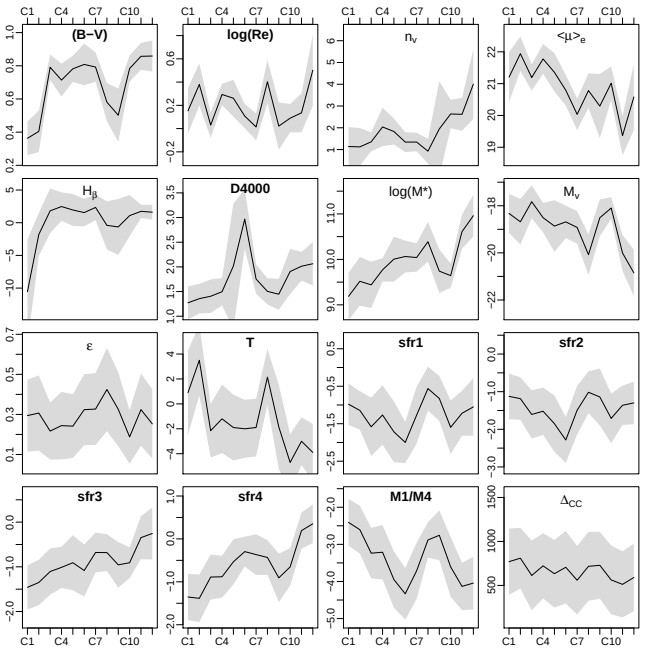

Fig. 8. Changes of the medians of the groups along the tree for the cluster sample. Shaded regions show 1σ dispersion around median.

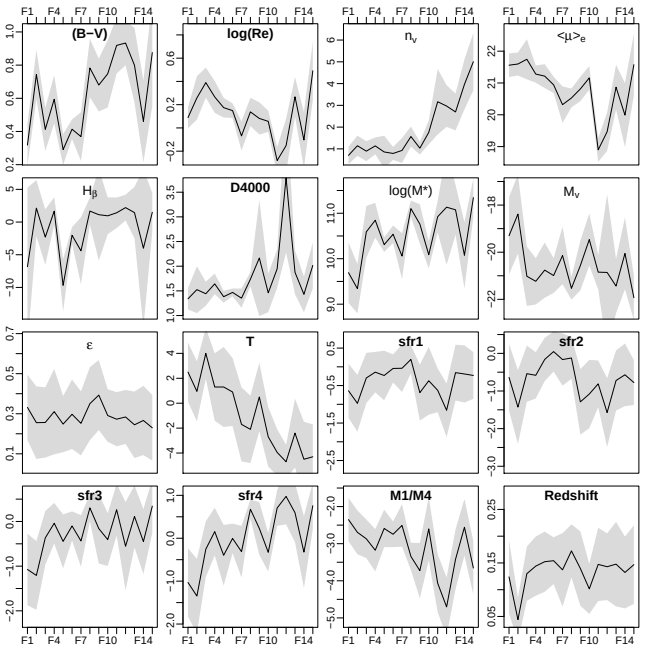

Fig. 9. Idem of Fig. 8 for the field sample.

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 8 and 9. 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 violinplots, the global evolution of the variables are similar in both samples. In particular the starting and ending points of the evolution curves are relatively close, meaning that the less and the most diversified groups are similar in both samples. There are however differences in the details, the evolution for the field sample seems more wiggling (that 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 evolutions can be then 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 continuous 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 (Sect. 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 8 and 9) 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. But 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: magnitude (V for us vs B), log(R_e), $\langle\mu\rangle_e$, H_β, and the mass (gas vs 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 evolutions 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 center, the intragroup variance being large (Figs. 4, 6a and 6b). This is apparently not in agreement with Valentinuzzi et al. (2011) that 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 8 and 9). The star formation rates ($sfr1$...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 except through harassment or strong interaction. 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 the 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 dimension 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 shows a reversal in its evolution. Fig. 6a 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, say 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) and Fraix-Burnet et al. (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 a stripped-off galaxies (D'Onofrio et al. 2015), but for C10 this is less clear. s So why C9 and C10 are 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, among the seven used to build the tree, H_β and log(R_e), do not seem to bring too much constraints (no obvious potentially strong trend). As a consequence the algorithm found the shortest path between the groups by optimizing the main trends for the maximum of parameters.

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, really more massive than C9, but as bluish as C9. This seems to be consistent with C8 having "just entered" into the cluster environment, possibly not yet stripped-off as C9. It turns out that C8 is one of the groups hosting jellyfish galaxies (Sect. 5.7).

What about the field sample? The global increase in mass is grossly present but there are several departures with a wiggling shape (Fig. 6b and Fig. 9). Contrarily to the cluster sample, only one parameter (n_V), among the seven ones used for the analysis, shows a regular trend although not perfectly monotonic. This means that the algorithm was not able to find a 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 that 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, 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 size of the cluster and field galaxies do not show any evolutionary trend in log(R_e) (Figs. 6a and 6b), 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 ellipticals, and much brighter, whatever their radius and their environment.

The expected evolution in radius with redshift is likely 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, in preparation).

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.

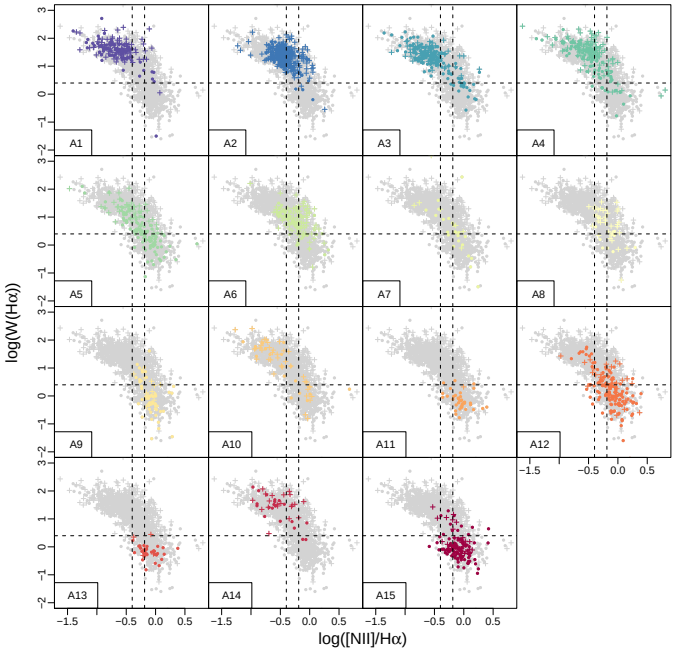

Fig. 10. Equivalent width of Hα vs [NII]/Hα . The vertical dotted lines corresponds to [NII]=/Hα=0.4 and 0.65 and the horizontal one is for W(Hα)=2.5 (see text). Dots are for cluster galaxies and crosses for field ones.

5.6. Emission line properties: gas deficit

Figures 10, 11, 12 show the behavior of the equivalent width of Hα vs. the ratio [NII]λ6584/Hα for the full, cluster and field samples respectively. The diagrams relate the W(Hα), a measure of the amount of gas per unit stellar mass, to the ratio [NII]λ6584/Hα (Cid Fernandes et al. 2010) which is a tracer of the ionisation mechanism. It cannot be $\gtrsim$ 0.65 for pure HII regions following the diagnostic diagrams of (Kauffmann et al. 2003). Typical HII nuclei have [NII]λ6584/Hα $\lesssim$ 0.4, LINERs and Seyfert-type AGN spectra have [NII]λ6584/Hα $\gtrsim$ 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 13 and 14). 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. The groups C1,2,8 and 9 mainly make the blue sequence of the cluster galaxies (see below). C3,4, 6, 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, such as C8, C9, most notably. However, C8 is predominantly made of late-type galaxies, C 9 seems to be the lower mass counterpart of C8. Both show high W(Hα) and hence large $sfr1$. It is interesting to recall that C6 to C9 and C10 to C12 define two distinct ensembles on the tree (Fig. 5) that 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 seen in the F groups, with F11 and F15 the only groups strongly

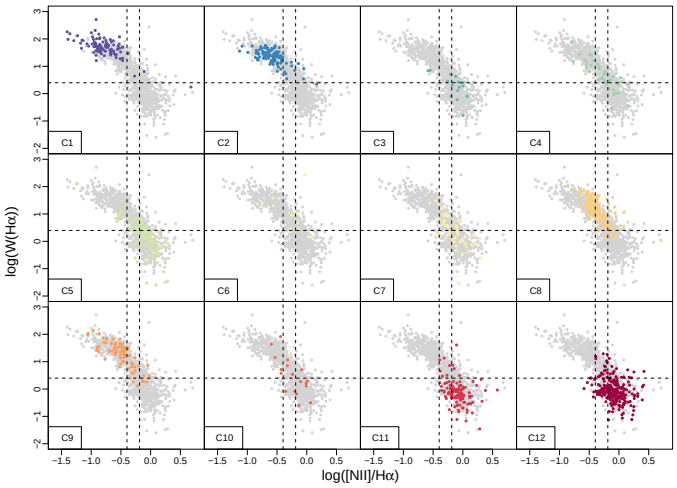

Fig. 11. Idem Fig. 10 for cluster galaxies.

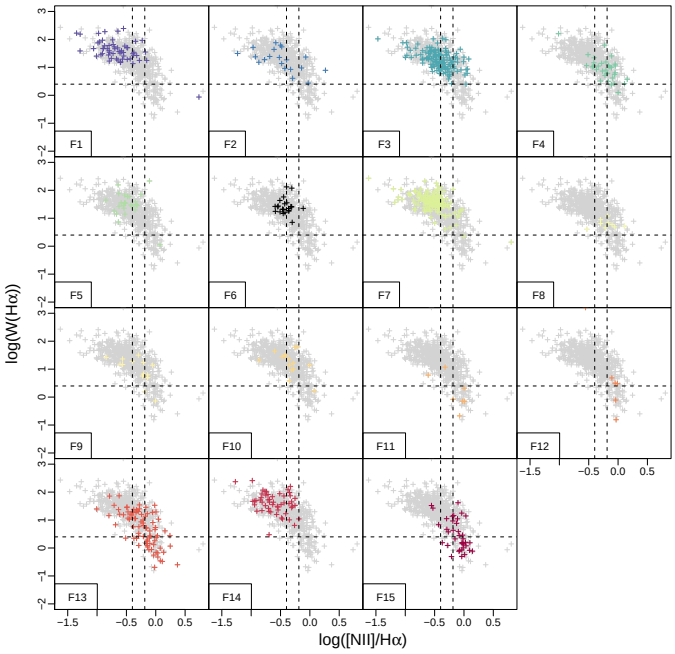

Fig. 12. Idem Fig. 10 for field galaxies.

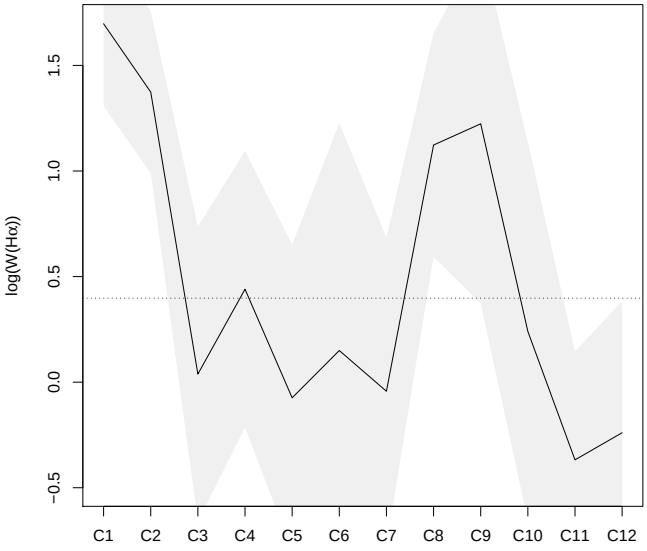

Fig. 13. Evolution of the median of W(Hα) for the cluster sample. Shaded regions show 1σ dispertion around median. The horizontal dotted line is for W(Hα)=2.5 (see text and Figs. 10, 11, 12).

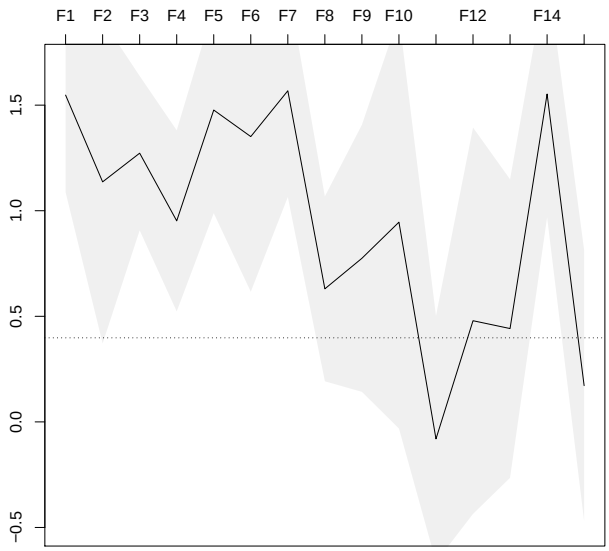

Fig. 14. Idem Fig. 13for the field sample.

departing from it. 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(Hα) among field galaxies is higher by a factor of a few than among cluster galaxies except for groups C8 and C9 (Figs 13 and 14). A large fraction of galaxies in the groups not dominated by HII nuclei falls below the limit of W(Hα)= 2.5 Å 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]λ6584/Hα associated with the red sequence of galaxies.

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 which show knotted arms of material resembling the tentacles of jellyfishes. 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 $\approx$ 10% in each group, and in F3, F9, and F13.

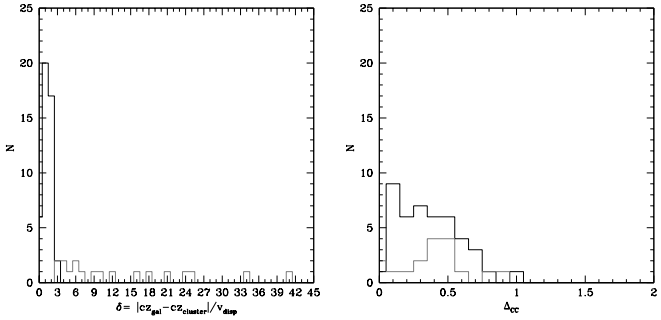

Fig. 15. Histograms for jellyfish galaxies that are in our cluster (black lines) or field (grey lines) samples, for the parameter δ (see text) and the distance Δ_{CC} to cluster center.

The groups F3 and F13 are large, and so the prevalence of jellyfish galaxies is very low, ≈2%. In F9 they are ≈10% of the sample. Among cluster group C2 is dominated by HII regions while C8 is shifted towards intermediate objects. Both groups are however predominantly above the limit of retired 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 of the cluster environment. Their clustercentric distance with respect to jellyfish galaxies classified as cluster member is only slightly skewed towards larger distances; the non-member distribution is well within the distribution of clustercentric 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 $cz_{\rm cluster}$ by less than 3 times the velocity dispersion of the cluster. Looking at the distribution of the variable $\delta = |cz - cz_{\rm cluster}|/\sigma$, we note that there are about 6 galaxies that are within $\delta \lesssim 6\sigma$ (Fig 15). Considering that jellyfish galaxies are probably high velocity intruders (Jaffé et al. 2018), these galaxies could be physically associated with the cluster. The remaining jellifishes 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 among field galaxies that are known to lead to tidal tails with star forming knots (e.g., Duc & Renaud 2013) loosely resembling the tails of cluster jellyfishes.

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. 16) 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 much influenced by the evolutionary history of the galaxies.

The shape of galaxies and their stellar population are therefore linked 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 among galactic components, the most famous being that between the black-hole mass and the bulge mass (Magorrian et al. 1998).

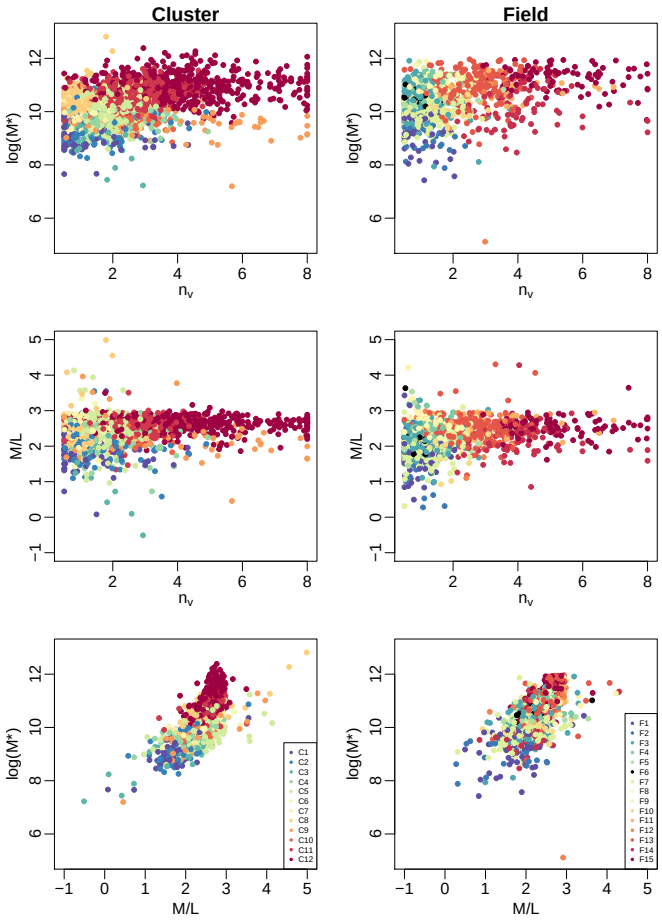

Fig. 16. Scatter plots between n_V, log(M*) and M/L for cluster (left column) and field (right column) galaxies.

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 case of catastrophic events such as major mergers, galaxies enter in a new configuration that rapidly forget the previous galaxy properties.

Co-evolution should be at the origin of many of the observed scaling relations.

5.9. The log(M*) vs log(R_e) relation

The log(R_e) vs log(M*) scatter plot shows that the galaxies in A15 and C12 are the only ones that define a clear Mass-Radius relation ((Fig. 17 and 18). 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. 19). The group distribution is highly discrim-

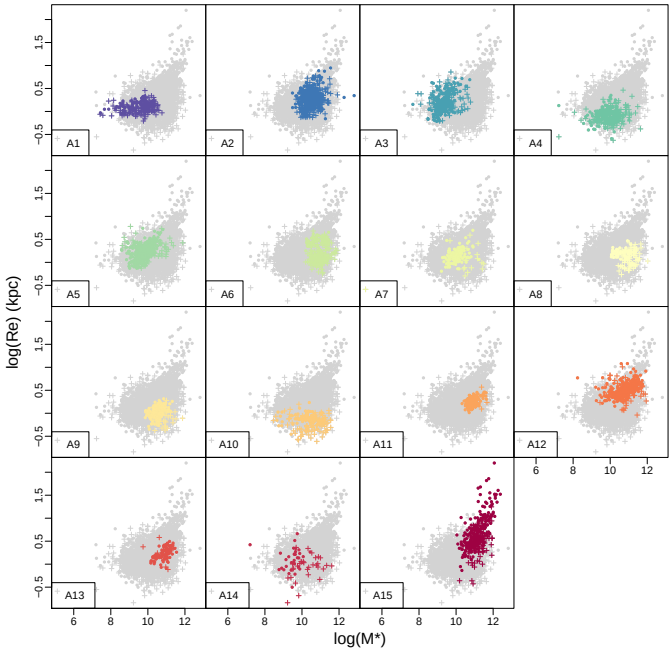

Fig. 17. Effective radius vs mass for the groups of the full sample. Gray points are the galaxies of the full sample.

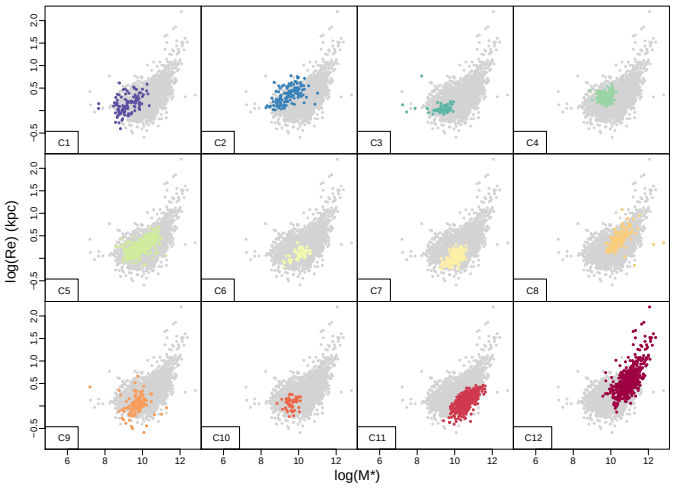

Fig. 18. Effective radius vs mass for the groups of the cluster sample. Gray points are the galaxies of the cluster sample only.

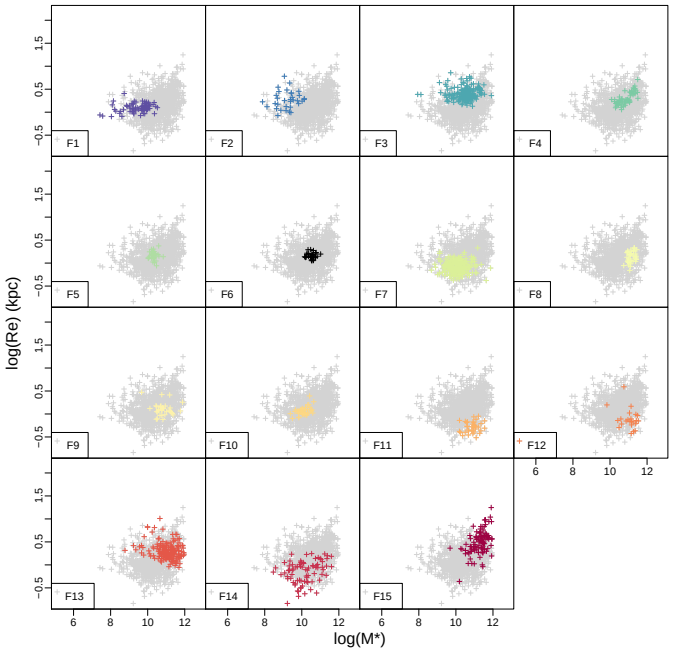

Fig. 19. Effective radius vs mass for the groups of the field sample. Gray points are the galaxies of the field sample only.

inated for the three samples, and is identical to the one found in Fraix-Burnet et al. (2010) and Fraix-Burnet et al. (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) unfortunately to unknown proportion 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 $\log(R_e)$ vs $\log(M^*)$ relation appears almost flat for galaxies with $M^* < 10^{10}$ $M_\odot$ both in ETGs and 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 relation $\log(R_e)$-$\log(M^*)$ is expected for the total mass M, that 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 by small size high mass ($\simeq 10^{11}$ $M_\odot$) elliptical galaxies, with an anomalous high D_{4000} and the highest $sfr4$. The number of 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), GAMA (Lange et al. 2015) and SAMI (Scott et al. 2017). However, super-dense galaxies are known to exist either in clusters (Valentinuzzi et al. 2010) and 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 around 4.4%. In our sample we have 24 members of the F12 groups over a total number of field galaxies of 1158 (2% frequency). If we add the F11 group, that is also made of quite massive and small objects, we reach a frequency of 5%. These value 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.

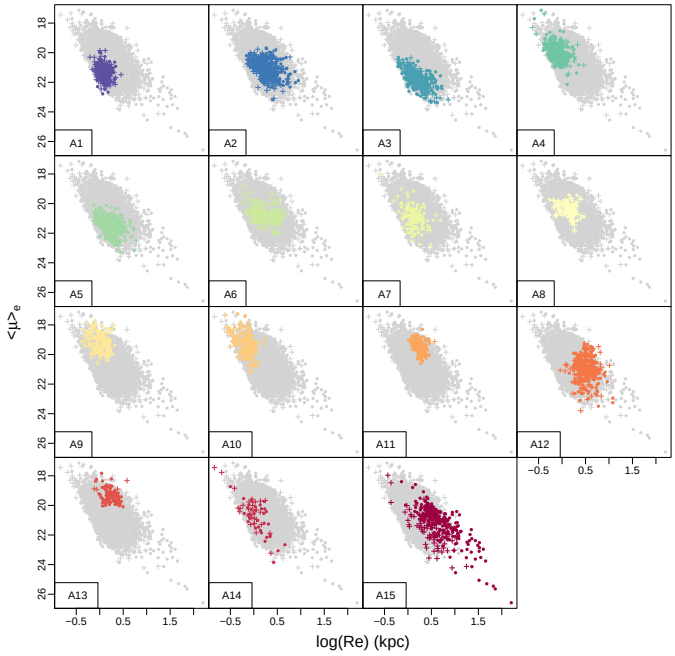

Fig. 20. Effective radius vs effective surface brightness for the groups of the full sample. Gray points are the galaxies of the full sample.

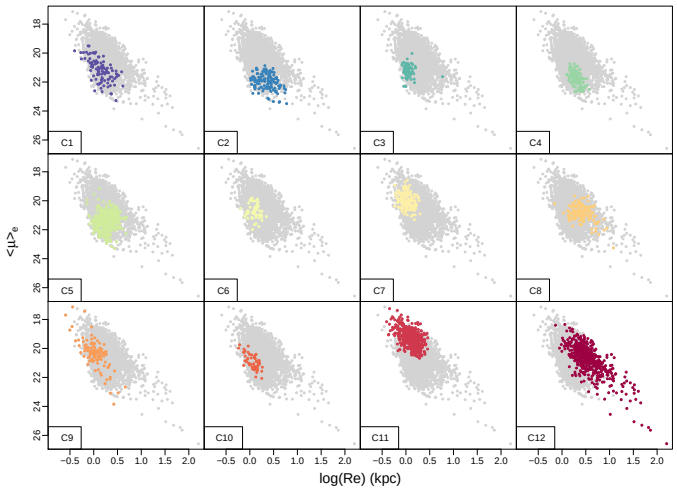

Fig. 21. Effective radius vs effective surface brightness for the groups of the cluster sample. Gray points are the galaxies of the cluster sample only.

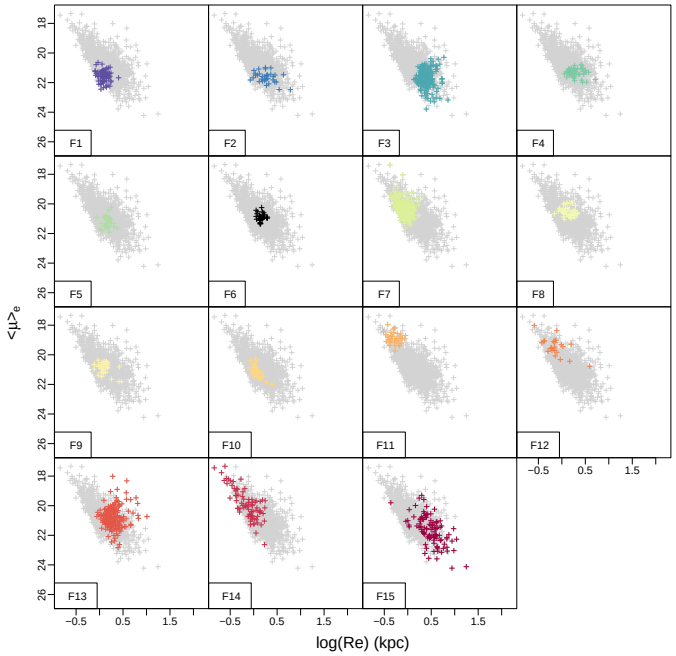

Fig. 22. Effective radius vs effective surface brightness for the groups of the field sample. Gray points are the galaxies of the field sample only.

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

The Kormendy relation $\langle\mu\rangle_e$ vs $\log(R_e)$ (Kormendy 1977) is visible on the three samples (Figs. 20, 21, 22) 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 $\log(R_e)$ vs $\log(M^*)$ plots (Sect. 5.9). The reason of this dispersion is the mixture of different populations, at different stages of evolution, quite alike 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).

It must be noticed that our groups pave the full sample distribution without much overlap. For instance, considering the Kormendy diagram for the cluster sample (Figs. 21), the regions to the top right are occupied by the three more 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 in two mass ranges, we find two parallel relations. Hence the Kormendy relation is not due to the mass nor to the environment, but to some co-evolution that depends on mass.

5.11. Color-magnitude diagram

The colour-magnitude diagram for the cluster galaxy sample (Fig. 23), shows two main regions that are occupied by groups: C1, 2 8 and 9, that corresponds to the blue sequence, and a longer and more populated region corresponding to the red sequence. The projection of the tree onto this diagram (Figs. C.1) depicts the evolutionary path: the evolution goes first 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 green valley. Hence, we have 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. 24) 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 of a large part of galaxies between the two diagrams. Even if this might partially be due to a Malmquist bias, it appears that the

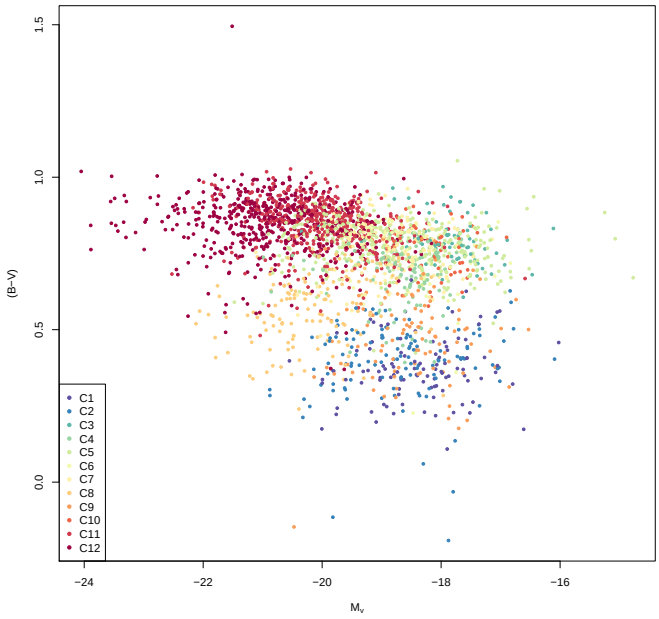

Fig. 23. Colour-magnitude diagram for the cluster sample.

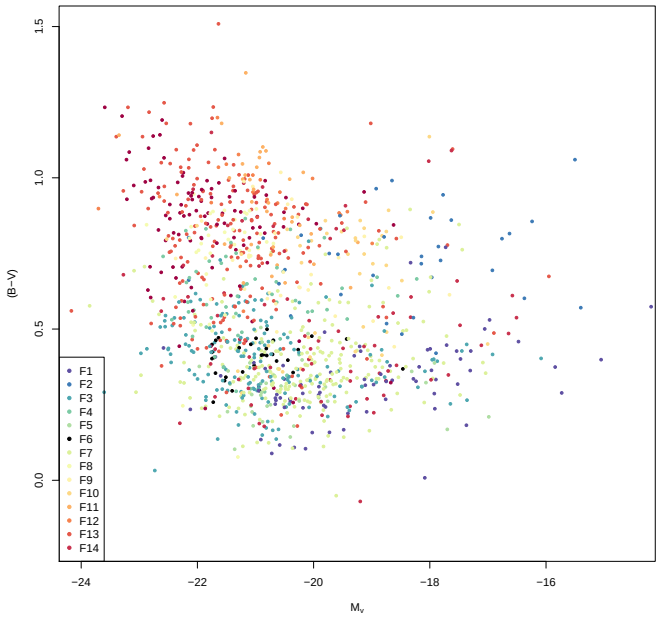

Fig. 24. Colour-magnitude diagram for the field sample.

beginning of the evolution from F1 to F2 is very similar to the case of the cluster sample (Figs. 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 Maximum Parsimony approach is able to reveal the characteristics of the evolution of different galaxy populations. In clusters the red and blue sequence are better defined and we can clearly identify the class of objects in the green valley, that are likely rejuvenated by episode of star formation or wet merger events. Note in particular the extended sequence of red and faint objects present in the cluster sample. They are almost absent in the field. These objects, once entered in the cluster environment rapidly loose 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 stellar formation has been hindered or suppressed.

6. Conclusions

This study aims at investigating the effect of cluster environment on the evolutionary scenario of galaxies. We have used a consequent sample of homogeneous data, from the WINGS catalogue, with about 2600 galaxies which are members of clusters, and about 1500 ones which are field galaxies.

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 pave 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 representativity and 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}, $\log(M^*)$. They are all fully documented, and we have checked that there is no redundancy nor strong correlation in this set.

We have obtained an evolutionary tree for the two samples 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 cluster sample, there is a distinct sub-structure, a distinct evolutionary path, that seems to gather rejunevated or stripped-off galaxies, and goes into the green valley (or even blue sequence) from the red sequence on the CMD 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 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" of 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 appearing 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 vs 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 identify known classes of objects by precising the properties, and to stick out 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. We have however shown that the evolutionary paths are somewhat complicated: even though galaxies globally become redder, more massive, less disky, there are many other possible trajectories, such as field galaxies which are bluer, more concentrated but more massive than cluster ones, or cluster galaxies that are stripped-off, etc. 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 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, Monthly Notices of the Royal Astronomical Society, 282, 1461
Blanco-Cuaresma, S. & Fraix-Burnet, D. 2018, Astronomy and Astrophysics - 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, Applied Statistics, 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., Bindoni, D., Fasano, G., et al. 2014, A&A, 572, A87
D'Onofrio, M., Marziani, P., & Buson, L. 2015, Frontiers in Astronomy and Space Sciences, 2, 4
D'Onofrio, M., Valentinuzzi, T., Fasano, G., et al. 2011, The Astrophysical Journal, 727, L6
Dressler, A., Lynden-Bell, D., Burstein, D., et al. 1987, ApJ, 313, 42
Duc, P.-A. & Renaud, F. 2013, in Lecture Notes in Physics, Berlin Springer Verlag, Vol. 861, Lecture Notes in Physics, Berlin Springer Verlag, ed. J. Souchay, S. Mathis, & T. Tokieda, 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, Cladistics: Perspectives on the reconstruction of evolutionary history, ed. T. Duncan & T. Stuessy (Columbia University Press, New York), 169–191
Fernández Lorenzo, M., Sulentic, J., Verdes-Montenegro, L., & Argudo-Fernández, M. 2013, MNRAS, 434, 325
Fraix-Burnet, D. 2011, Monthly Notices of the Royal Astronomical Society: Letters, 416, L36
Fraix-Burnet, D. 2016, EAS Publications Series, Vol. 77, Statistics for Astrophysics: Clustering and Classification, ed. D. Fraix-Burnet & S. Girard (EDP Sciences), 221–257
Fraix-Burnet, D., Chattopadhyay, T., Chattopadhyay, A. K., Davoust, E., & Thuillard, M. 2012, Astronomy and Astrophysics, 545, A80
Fraix-Burnet, D., Choler, P., Douzery, E., & Verhamme, A. 2006a, Journal of Classification, 23, 31
Fraix-Burnet, D., Douzery, E., Choler, P., & Verhamme, A. 2006b, Journal of Classification, 23, 57
Fraix-Burnet, D., Dugué, M., Chattopadhyay, T., Chattopadhyay, A. K., & Davoust, E. 2010, MNRAS, 407, 2207
Fraix-Burnet, D., Thuillard, M., & Chattopadhyay, A. K. 2015, Frontiers in Astronomy and Space Sciences, 2 [http://arxiv.org/abs/1508.06756]
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, Molecular Biology and Evolution, 23, 1997
Genel, S., Vogelsberger, M., Springel, V., et al. 2014, MNRAS, 445, 175
Hendy, M. & Penny, D. 1982, Mathematical Biosciences, 59, 277
Hennig, W. 1965, Annual Review of Entomology, 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, Monthly Notices of the Royal Astronomical Society, 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, Systematic Zoology, 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
Murtagh, F. & Legendre, P. 2014, Journal of Classification, 31, 274–295
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, Molecular Biology and Evolution, 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)
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

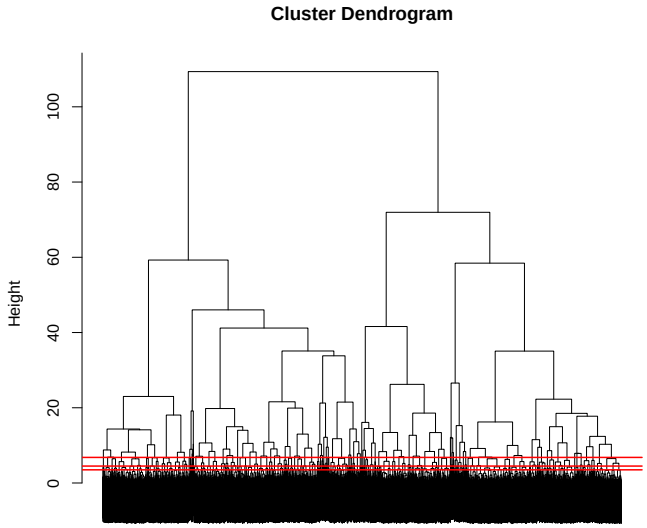

Fig. A.1. Dendrogram of the hierarchical clustering algorithm for the full sample. The horizontal red lines indicate from top to bottom the cuts defining the 100, 200 and 300 pre-clusters respectively.

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 Sect. 3.2 and Sect. 3.3 how the number of pre-clusters and the contours of the phylogenetic groups have been 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 violinplots (Fig. A.3 and A.4) confirms 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 tend 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.

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 violinplots (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 clearcut separation.

The sequence of the changes of the variables along the tree (Fig. A.6) should be read with caution, the two branch must be seen as two parallel separate evolutionary paths. The striking behaviour is the decreasing colour, that is galaxies become bluer, especially in the 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 the branch A8 to A15 than in A7, but is slightly lower in the branch A6 to A2. The morphology also in the second branch clearly evolves towards spiral types. We think that the changes within each branch is 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 have not changed significantly our result, and it is doubtful that the evolution (and the diversification) of galaxies would depend on only one single property.

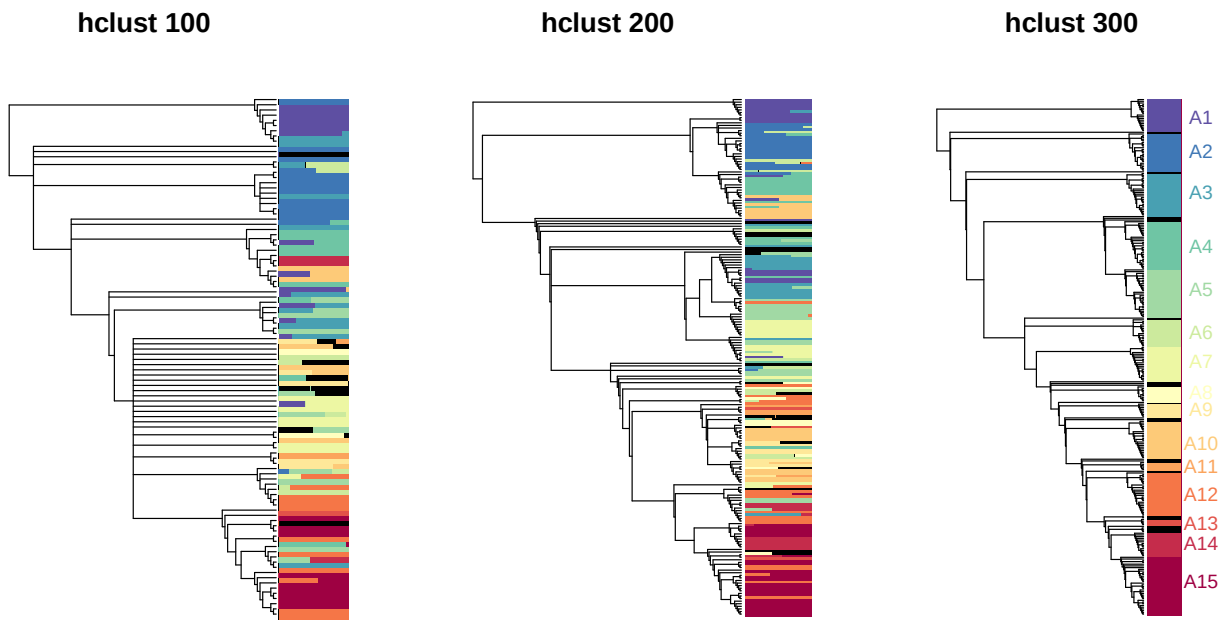

Fig. A.2. Trees obtained for the full sample with 100 (left), 200 (middle) and 300 (right) pre-clusters (see Sect. 3.2). The tree on the right is the one shown in Fig. 3. The coloured bars at each tip of the trees give the distribution of the galaxies of the corresponding pre-cluster in the groups defined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

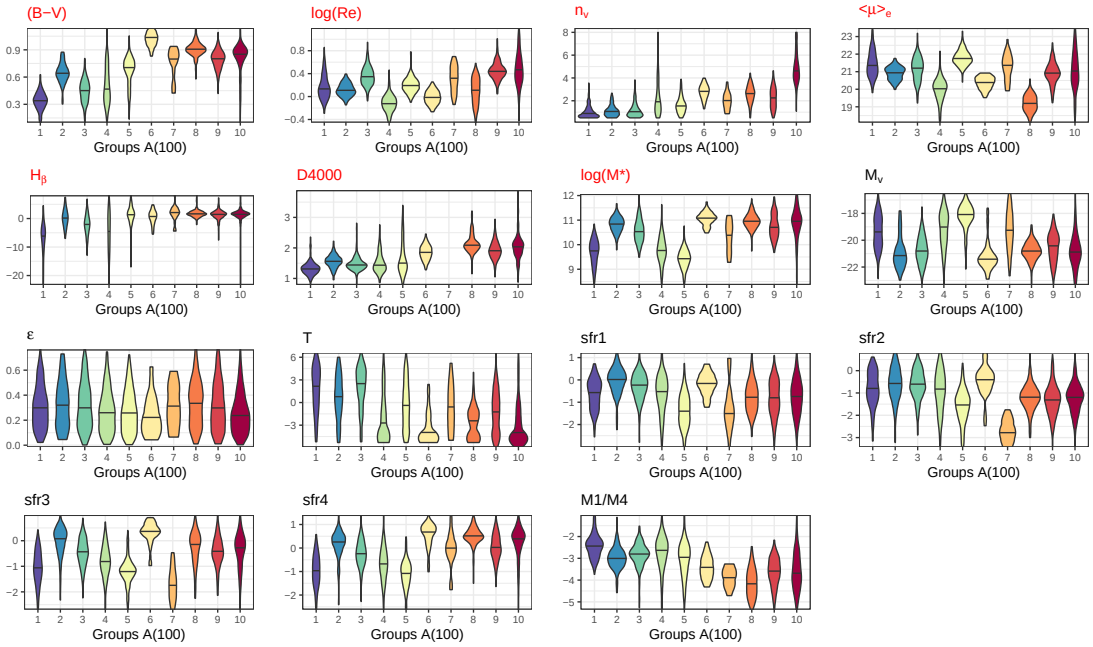

Fig. A.3. Violinplots for groups defined on the tree with 100 pre-clusters (left in Fig. A.2) in the same way as described in Sect. 3.3. This figure should be compared with Fig. 4 and Fig. A.4. The violinplot of group 7 for parameter D_{4000} spreads between 5 and 9 and is thus well above the diagram.

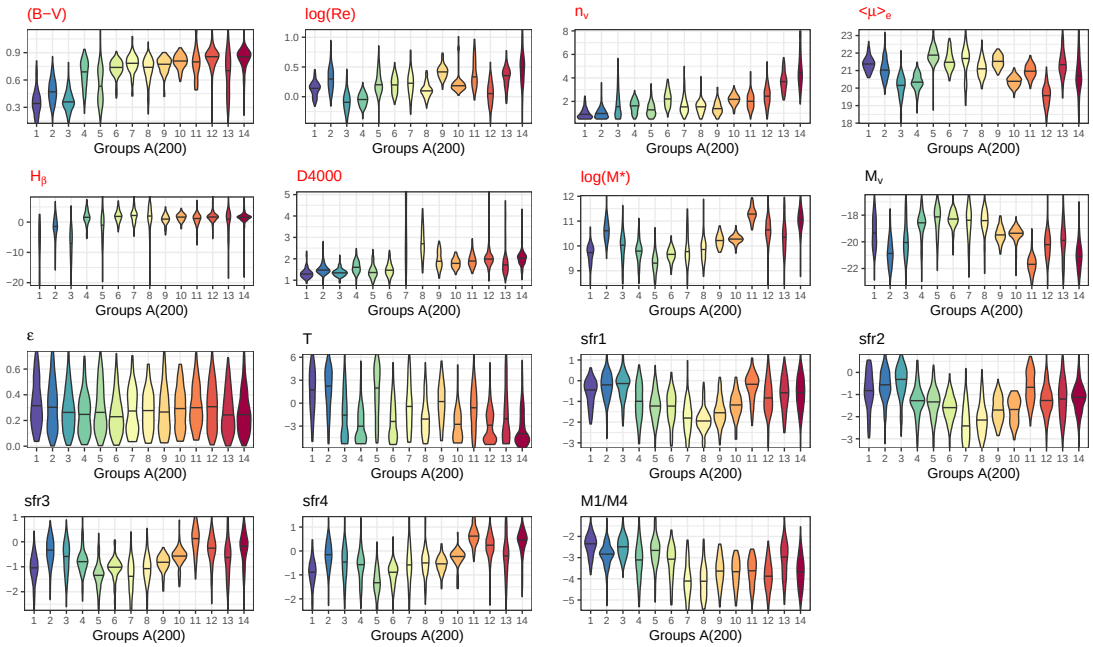

Fig. A.4. Violinplots for groups defined on the tree with 200 pre-clusters (middle in Fig. A.2) in the same way as described in Sect. 3.3. This figure should be compared with Fig. 4 and Fig. A.3.

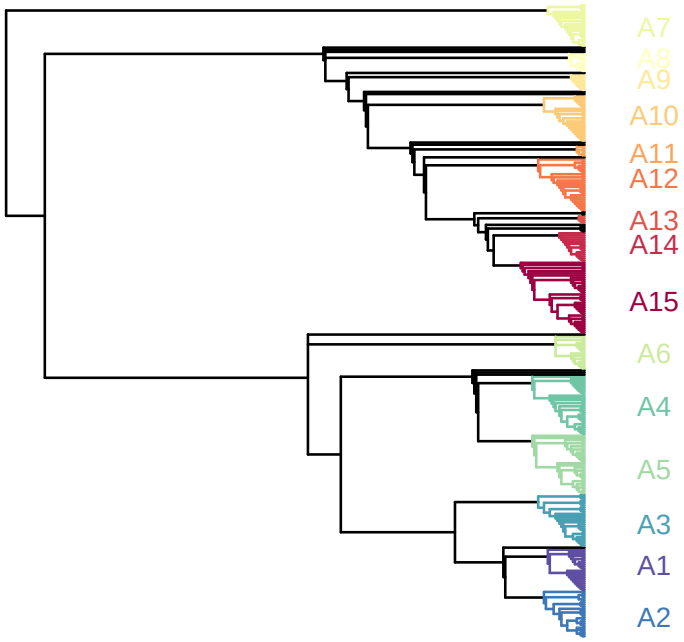

Fig. A.5. Same tree as in Fig. 3 for the full sample but rooted with group A7. Colours and names of groups are identical.

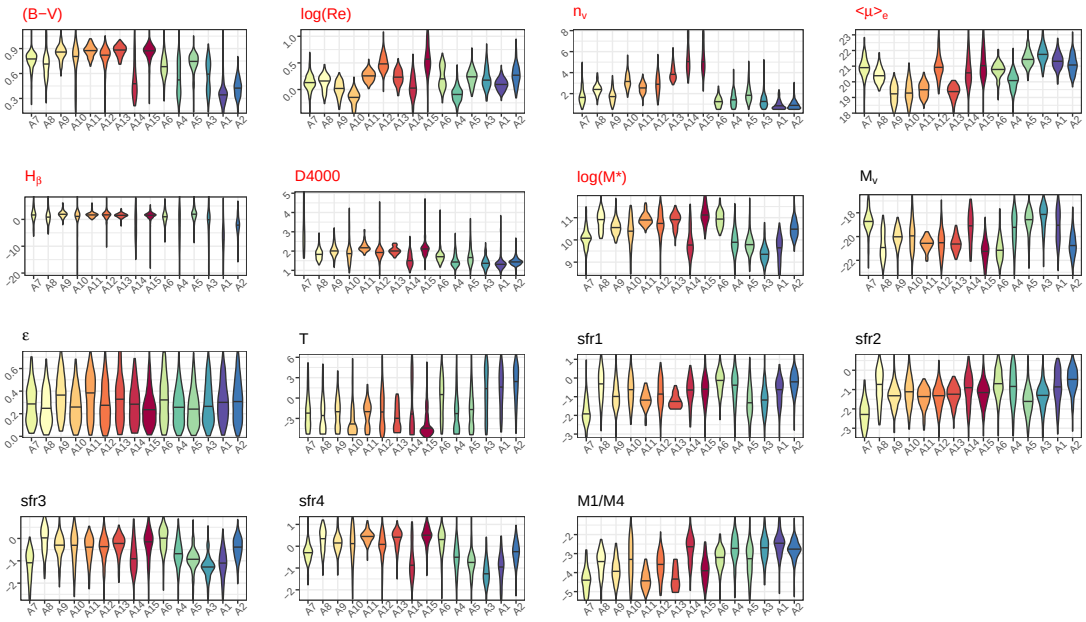

Fig. A.6. Same plot as in Fig. 4 for the full sample but rooted with group A7 according to tree in Fig. A.5. Colours and names of groups are identical..

Appendix B: Statistical correlations among the variables

These figures are described in Sect. 5.1.

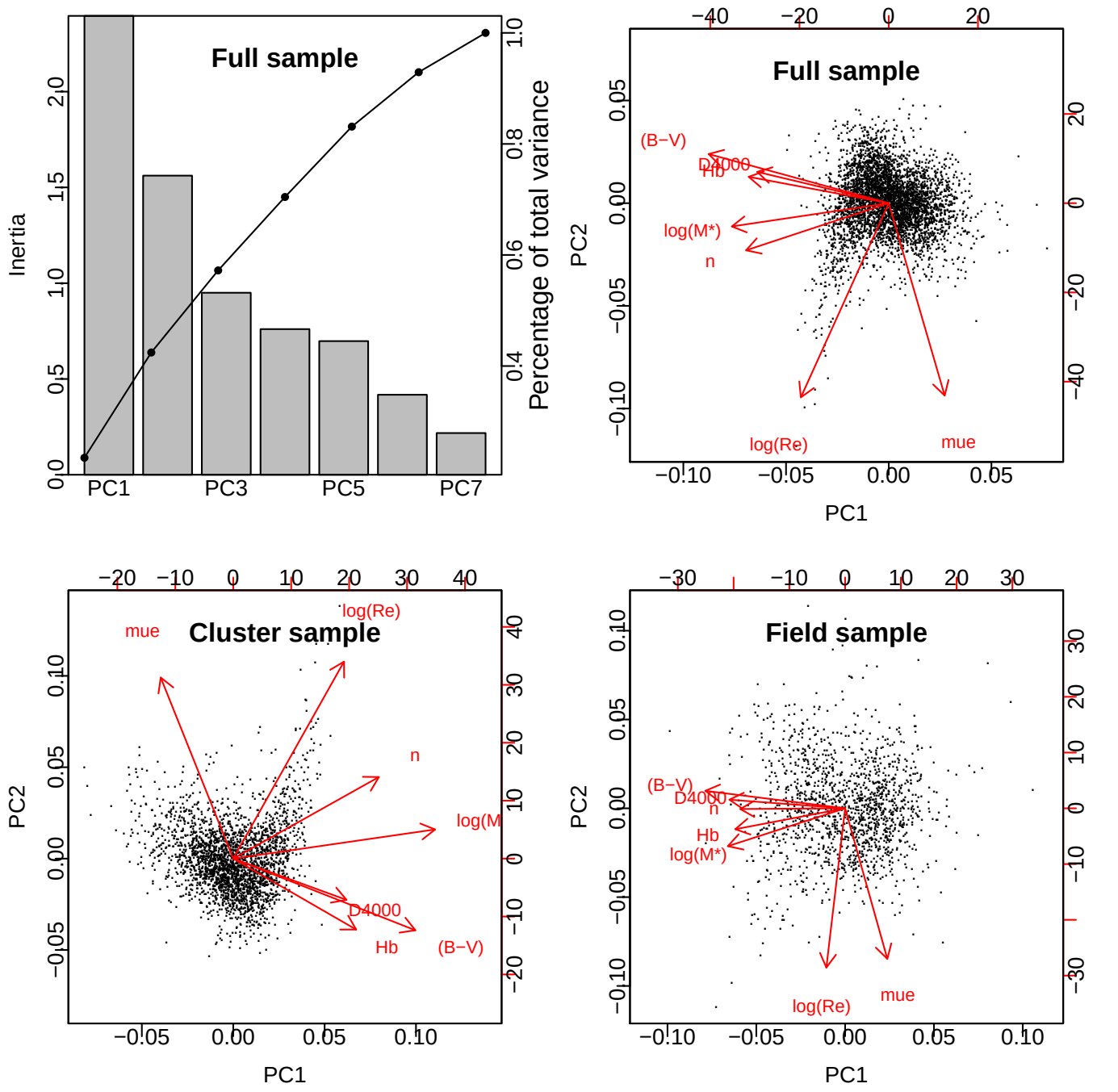

Fig. B.1. Principal Component Analysis using the seven parameters used for the clustering and MP analyses (Sect. 5.1). *(Top left)* Screeplot for the full sample showing the variance (eigenvalues) of the seven Principal Components (PCs). The curve represents the cumulative percentage of the total variance. The screeplot for the cluster and field sample are very similar. *(Top right)* Projection of the data on the plane of the first two PCs with arrow showing the loadings and direction of the parameters. The same diagram are plotted for the cluster (*bottom left*) and field (*bottom right*) samples.

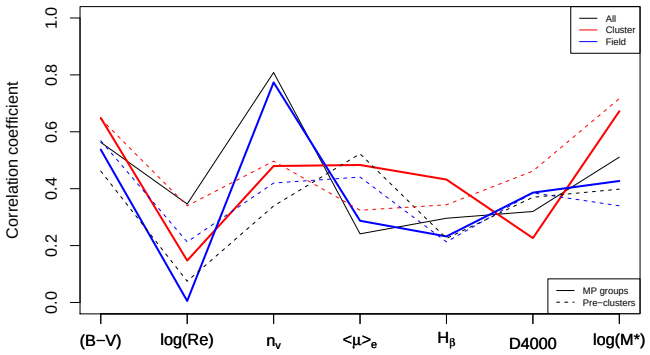

Fig. B.2. Correlation (absolute value) of the seven variables with the groups (solid lines) and the pre-clusters (dashed lines; Sect. 5.1).

Appendix C: Projection of the tree on the CMD diagram

These figures are described in Sect. 5.11. The trees in Fig. 23 and 24 are projected on the bivariate plot B–V vs M_V. The tips of the trees are the medians of these parameters for each group. The internal nodes are given values by minimizing the squared changes of the parameters along the branches (squared-change parsimony ancestral state reconstruction, Maddison 1991).

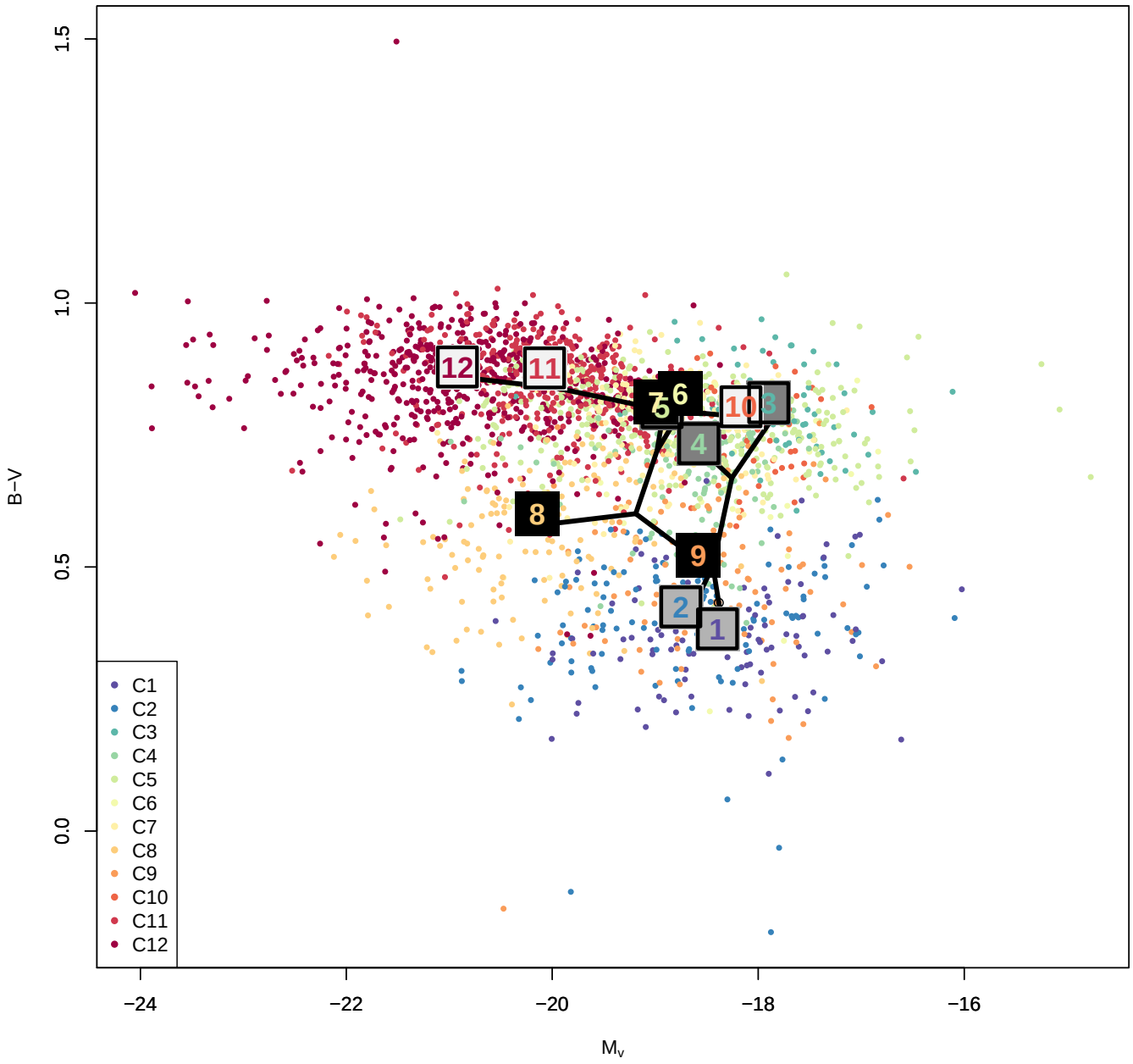

Fig. C.1. Tree projected onto the CMD diagram for the cluster sample.

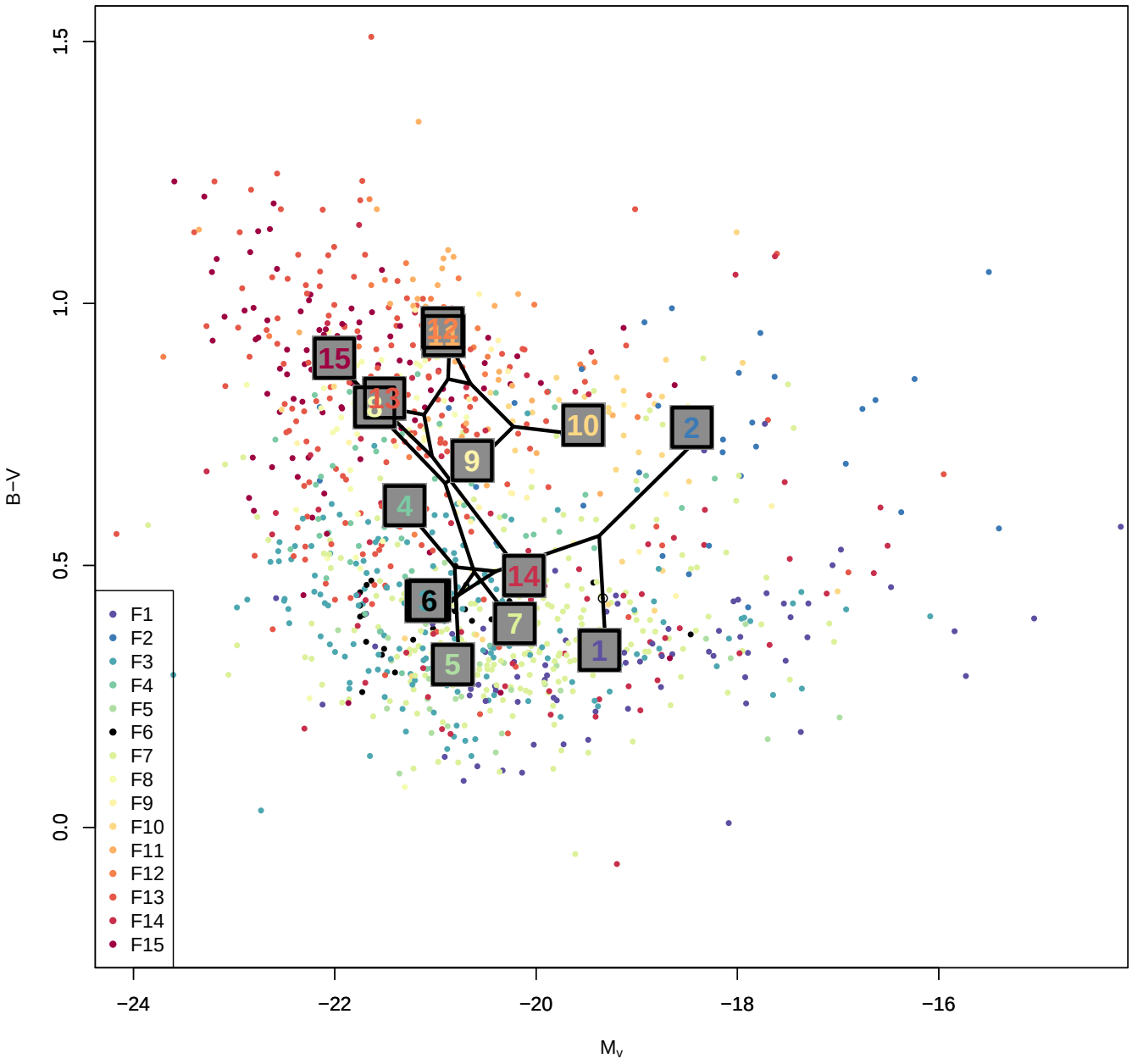

Fig. C.2. Tree projected onto the CMD diagram for the field sample. The median values of the groups F3 and F6 are superimposed on this plot.