Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **457,** 4360–4383 (2016)
Advance Access publication 2016 January 27

doi:10.1093/mnras/stw221

Mapping stellar content to dark matter haloes – II. Halo mass is the main driver of galaxy quenching

Ying Zu★ and Rachel Mandelbaum

McWilliams Center for Cosmology, Department of Physics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA

Accepted 2016 January 25. Received 2015 December 29; in original form 2015 September 22

ABSTRACT

We develop a simple yet comprehensive method to distinguish the underlying drivers of galaxy quenching, using the clustering and galaxy–galaxy lensing of red and blue galaxies in Sloan Digital Sky Survey. Building on the iHOD framework developed by Zu & Mandelbaum, we consider two quenching scenarios: (1) a *'halo'* quenching model in which halo mass is the sole driver for turning off star formation in both centrals and satellites; and (2) a *'hybrid'* quenching model in which the quenched fraction of galaxies depends on their stellar mass, while the satellite quenching has an extra dependence on halo mass. The two best-fitting models describe the red galaxy clustering and lensing equally well, but halo quenching provides significantly better fits to the blue galaxies above $10^{11}\,h^{-2}\,\mathrm{M}_{\odot}$. The halo quenching model also correctly predicts the average halo mass of the red and blue centrals, showing excellent agreement with the direct weak lensing measurements of locally brightest galaxies. Models in which quenching is not tied to halo mass, including an age-matching model in which galaxy colour depends on halo age at fixed M_*, fail to reproduce the observed halo mass for massive blue centrals. We find similar critical halo masses responsible for the quenching of centrals and satellites ($\sim 1.5 \times 10^{12}\,h^{-1}\,\mathrm{M}_{\odot}$), hinting at a uniform quenching mechanism for both, e.g. the virial shock heating of infalling gas. The success of the iHOD halo quenching model provides strong evidence that the physical mechanism that quenches star formation in galaxies is tied principally to the masses of their dark matter haloes rather than the properties of their stellar components.

Key words: gravitational lensing: weak – methods: statistical – galaxies: luminosity function, mass function – cosmology: observations – large-scale structure of Universe.

1 INTRODUCTION

The quenching of galaxies, namely, the relatively abrupt shutdown of star formation activities, gives rise to two distinctive populations of quiescent and active galaxies, most notably manifested in the strong bimodality of galaxy colours (Strateva et al. 2001; Baldry et al. 2006). The underlying driver of quenching, whether it be stellar mass, halo mass, or environment, should produce an equally distinct split in the spatial clustering and weak gravitational lensing between the red and blue galaxies. Recently, Zu & Mandelbaum (2015, hereafter Paper I) developed a powerful statistical framework, called the iHOD model, to interpret the spatial clustering (i.e. the projected galaxy autocorrelation function $w_{\rm p}$) and the galaxy–galaxy (g–g) lensing (i.e. the projected surface density contrast $\Delta\Sigma$) of the overall galaxy population in the Sloan Digital Sky Survey (SDSS; York et al. 2000), while establishing a robust mapping between the observed distribution of stellar mass to that of the underlying dark matter haloes. In this paper, by introducing two empirically motivated and physically meaningful quenching models within iHOD, we hope to robustly identify the dominant driver of galaxy quenching, while providing a self-consistent framework to explain the bimodality in the spatial distribution of galaxies.

Galaxies cease to form new stars and become quenched when there is no cold gas. Any physical process responsible for quenching has to operate in one of three following modes: (1) it heats up the gas to high temperatures and stops hot gas from cooling efficiently (e.g. gravitational collapse and various baryonic feedback; see Benson 2010, for a review); (2) it depletes the cold gas reservoir via secular stellar mass growth or sudden removal by external forces (e.g. tidal and ram pressure; Gunn & Gott 1972); and (3) it turns off gas supply by slowly shutting down accretion (e.g. strangulation; Balogh & Morris 2000). However, due to the enormous complexity in the formation history of individual galaxies, multiple quenching modes may play a role in the history of quiescent galaxies. Therefore, it is more promising to focus on the underlying physical driver of the

★ E-mail: yzu@cmu.edu

average quenching process, which is eventually tied to either the dark matter mass of the host haloes, the galaxy stellar mass, or the small/large-scale environment density that the galaxies reside in, hence the so-called halo, stellar mass, and environment quenching mechanisms, respectively.

Halo quenching has provided one of the most coherent quenching scenarios from the theoretical perspective. In haloes above some critical mass ($M_{\rm shock} \sim 10^{12}\,h^{-1}\,{\rm M}_{\odot}$), virial shocks heat gas inflows from the intergalactic medium, preventing the accreted gas from directly fuelling star formation (Binney 1977, 2004; Birnboim & Dekel 2003; Katz et al. 2003; Kereš et al. 2005, 2009). Additional heating from, e.g. the active galactic nuclei (AGNs) then maintains the gas coronae at high temperature (Croton et al. 2006). For haloes with $M_{\rm h} < M_{\rm shock}$, the incoming gas is never heated to the virial temperature due to rapid post-shock cooling, therefore penetrating the virial boundary into inner haloes as cold flows. This picture, featuring a sharp switch from the efficient stellar mass buildup via filamentary cold flow into low-mass haloes, to the halt of star formation due to quasi-spherical hot-mode accretion in haloes above $M_{\rm shock}$, naturally explains the colour bimodality, particularly the paucity of galaxies transitioning from blue, star-forming galaxies to the red sequence of quiescent galaxies (Cattaneo et al. 2006; Dekel & Birnboim 2006). To first order, halo quenching does not discriminate between centrals and satellites, as both are immersed in the same hot gas coronae that inhibits star formation. However, since the satellites generally lived in lower mass haloes before their accretion and may have retained some cold gas after accretion, the dependence of satellite quenching on halo mass should have a softer transition across $M_{\rm shock}$, unless the quenching by hot haloes is instantaneous.

Observationally, by studying the dependence of the red galaxy fraction $f^{\rm red}$ on stellar mass M_* and galaxy environment $\delta_{\rm 5NN}$ (i.e. using distance to the fifth nearest neighbour) in both the SDSS and zCOSMOS, Peng et al. (2010, hereafter P10) found that $f^{\rm red}$ can be empirically described by the product of two independent trends with M_* and $\delta_{\rm 5NN}$, suggesting that stellar mass and environment quenching are at play. By using a group catalogue constructed from the SDSS spectroscopic sample, Peng et al. (2012) further argued that, while the stellar mass quenching is ubiquitous in both centrals and satellites, environment quenching mainly applies to the satellite galaxies.

However, despite the empirically robust trends revealed in P10, the interpretations for both the stellar mass and environment trends are obscured by the complex relation between the two observables and other physical quantities. In particular, since the observed M_* of central galaxies is tightly correlated with halo mass $M_{\rm h}$ (with a scatter $\sim$0.22 dex; see Paper I), a stellar mass trend of $f^{\rm red}$ is almost indistinguishable with an underlying trend with halo mass. By examining the interrelation among M_*, $M_{\rm h}$, and $\delta_{\rm 5NN}$, Woo et al. (2013) found that the quenched fraction is more strongly correlated with $M_{\rm h}$ at fixed M_* than with M_* at $M_{\rm h}$, and the satellite quenching by $\delta_{\rm 5NN}$ can be re-interpreted as halo quenching by taking into account the dependence of quenched fraction on the distances to the halo centres. The halo quenching interpretation of the stellar and environment quenching trends is further demonstrated by Gabor & Davé (2015), who implemented halo quenching in cosmological hydrodynamic simulations by triggering quenching in regions dominated by hot ($10^{5.4}$ K) gas. They reproduced a broad range of empirical trends detected in P10 and Woo et al. (2013), suggesting that the halo mass remains the determining factor in the quenching of low-redshift galaxies.

Another alternative quenching model is the so-called age-matching prescription of Hearin & Watson (2013) and its recently updated version of Hearin et al. (2014). Age-matching is an extension of the 'subhalo abundance matching' (SHAM; Conroy, Wechsler & Kravtsov 2006) technique, which assigns stellar masses to individual subhaloes (including both main and subhaloes) in the N-body simulations based on halo properties like the peak circular velocity (Reddick et al. 2013). In practice, after assigning M_* using SHAM, the age-matching method further matches the colours of galaxies at fixed M_* to the ages of their matched haloes, so that older haloes host redder galaxies. In essence, the age-matching prescription effectively assumes a stellar mass quenching, as the colour assignment is done at fixed M_* regardless of halo mass or environment, with a secondary quenching via halo formation time. Therefore, the age-matching quenching is very similar to the M_*-dominated quenching of P10, except that the second variable is halo formation time rather than galaxy environment.

The key difference between the $M_{\rm h}$- and M_*-dominated quenching scenarios lies in the way central galaxies become quiescent. One relies on the stellar mass, while the other on the mass of the host haloes, producing two very different sets of colour-segregated stellar-to-halo mass relations (SHMRs). At fixed halo mass, if stellar mass quenching dominates, the red centrals should have a higher average stellar mass than the blue centrals; in the halo quenching scenario the two coloured populations at fixed halo mass would have similar average stellar masses, but there is still a trend for massive galaxies to be red because higher mass haloes host more massive galaxies. This difference in SHMRs directly translates to two distinctive ways the red and blue galaxies populate the underlying dark matter haloes according to their M_* and $M_{\rm h}$, hence two different spatial distributions of galaxy colours.

Therefore, by comparing the $w_{\rm p}$ and $\Delta\Sigma$ predicted from each quenching model to the measurements from SDSS, we expect to robustly distinguish the two quenching scenarios. The `iHOD` framework we developed in Paper I is ideally suited for this task. The `iHOD` is a global 'halo occupation distribution' (HOD) model defined on a 2D grid of M_* and $M_{\rm h}$, which is crucial to modelling the segregation of red and blue galaxies in their M_* distributions at fixed $M_{\rm h}$. The `iHOD` quenching constraint is fundamentally different and ultimately more meaningful compared to approaches in which colour-segregated populations are treated independently (e.g. Tinker et al. 2013; Rodríguez-Puebla et al. 2015). Our `iHOD` quenching model automatically fulfils the consistency relation which requires that the sum of red and blue SHMRs is mathematically identical to the overall SHMR. More importantly, the `iHOD` quenching model employs only four additional parameters that are directly related to the average galaxy quenching, while most of the traditional approaches require $\sim$20 additional parameters, rendering the interpretation of constraints difficult. Furthermore, the `iHOD` framework allows us to include $\sim$80 per cent more galaxies than the traditional HODs and take into account the incompleteness of stellar mass samples in a self-consistent manner.

This paper is organized as follows. We describe the selection of red and blue samples in Section 2. In Section 3, we introduce the parametrizations of the two quenching models and derive the `iHOD`s for each colour. We also briefly describe the signal measurement and model prediction in Sections 2 and 3, respectively, but refer readers to Paper I for more details. The constraints from both quenching mode analyses are presented in Section 4. We perform a thorough model comparison using two independent criteria in Section 5 and discover that halo quenching model is strongly favoured by the

data. In Section 6, we discuss the physical implications of the halo quenching model and compare it to other works in Section 7. We conclude by summarizing our key findings in Section 8.

Throughout this paper and Paper I, we assume a ΛCDM cosmology with $(\Omega_{\rm m}, \Omega_\Lambda, \sigma_8, h) = (0.26, 0.74, 0.77, 0.72)$. All the length and mass units in this paper are scaled as if the Hubble constant were $100\,{\rm km\,s^{-1}Mpc^{-1}}$. In particular, all the separations are comoving distances in units of either h^{-1} kpc or h^{-1} Mpc, and the stellar mass and halo mass are in units of $h^{-2}\,{\rm M}_\odot$ and $h^{-1}\,{\rm M}_\odot$, respectively. Unless otherwise noted, the halo mass is defined by $M_{\rm h} \equiv M_{200m} = 200\bar{\rho}_m(4\pi/3)r^3_{200m}$, where r_{200m} is the corresponding halo radius within which the average density of the enclosed mass is 200 times the mean matter density of the Universe, $\bar{\rho}_m$. For the sake of simplicity, $\ln x = \log_e x$ is used for the natural logarithm, and $\lg x = \log_{10} x$ is used for the base-10 logarithm.

2 SAMPLE SELECTION AND SIGNAL MEASUREMENT

In this section we describe the SDSS data used in this paper, especially the selection of the red and blue galaxies within the stellar mass samples, and the measurements of the galaxy clustering and the g–g lensing signals. We briefly describe the overall large-scale structure sample and the signal measurement, same as that used in Paper I, below in Sections 2.1 and 2.3, respectively, and refer readers to Paper I for details. Here, we focus more on the colour cut we employ to divide the galaxies into red and blue populations in Section 2.2.

2.1 NYU–VAGC and stellar mass samples

We make use of the final data release of the SDSS (DR7; Abazajian et al. 2009), which contains the completed data set of the SDSS-I and the SDSS-II. In particular, we obtain the main galaxy sample (MGS) data from the `dr72` large-scale structure sample `bright0` of the 'New York University Value Added Catalogue' (NYU–VAGC), constructed as described in Blanton et al. (2005). The `bright0` sample includes galaxies with $10 < m_r < 17.6$, where m_r is the r-band Petrosian apparent magnitude, corrected for Galactic extinction. We apply the 'nearest-neighbour' scheme to correct for the 7 per cent galaxies that are without redshift due to fibre collision, and use data exclusively within the contiguous area in the North Galactic Cap and regions with angular completeness greater than 0.8. The final sample used for the galaxy clustering analysis includes 513 150 galaxies over a sky area of 6395.49 deg^2. A further 5 per cent of the area is eliminated for the lensing analysis, due to the absence of source galaxies in that area.

As discussed in Paper I, we further restrict our analysis to galaxies above a 'mixture limit', defined as the stellar mass threshold above which the galaxy sample is relatively complete with a fair mix of red and blue galaxies. The functional form we adopt to describe the mixture limit $M_*^{\rm mix}(z)$ is

$$\lg\left(\frac{M_*^{\rm mix}}{h^{-2}\,{\rm M}_\odot}\right) = 5.4 \times (z - 0.025)^{0.33} + 8.0, \quad (1)$$

shown as the thick yellow curves in Fig. 3 (discussed further below). By taking into account the sample incompleteness in a self-consistent way, the `iHOD` model is able to model the lensing and clustering statistics of all galaxies above the mixture limit, ∼84 per cent more than the traditional HOD models typically include from the same catalogue.

We employ the stellar mass estimates from the latest MPA/JHU value-added galaxy catalogue.[1] The stellar masses were estimated based on fits to the SDSS photometry following the philosophy of Kauffmann et al. (2003) and Salim et al. (2007), and assuming the Chabrier (Chabrier 2003) initial mass function and the Bruzual & Charlot (2003) stellar population synthesis model. The MPA/JHU stellar mass catalogue is then matched to the NYU-VAGC `bright0` sample. We identify valid, unambiguous MPA/JHU stellar mass estimates for all but 32 327 (6.3 per cent) of the MGS galaxies. For those unmatched galaxies, we predict their stellar masses using the overall scaling between the two stellar mass estimates, depending on the $g - r$ colours (k-corrected to $z = 0.1$).

As sources for the g–g lensing measurement, we use a catalogue of background galaxies (Reyes et al. 2012) with a number density of 1.2 arcmin^{-2} with weak lensing shears estimated using the re-Gaussianization method (Hirata & Seljak 2003) and photometric redshifts from Zurich Extragalactic Bayesian Redshift Analyzer (Feldmann et al. 2006). The catalogue was characterized in several papers that describe the data, and use both the data and simulations to estimate systematic errors (see Mandelbaum et al. 2012, 2013; Nakajima et al. 2012; Reyes et al. 2012).

2.2 Separating sample into red and blue

We define quenching by the $(g - r)^{0.1}$ colour (after K-correction to $z = 0.1$) for three reasons: (1) colour bimodality is very stable across different environments and redshifts (Baldry et al. 2006); (2) observationally colour is very easy to measure robustly, without the need to fit galaxy morphology or brightness profile; and (3) physically colour is the result of integrated star formation history, largely immune to incidental or minor star formation episodes. In addition, we aim to model and compare to the two separable quenching trends with stellar mass and environment that revealed in P10, who originally chose optical colour as the quenching indicator.

Fig. 1 illustrates the colour–stellar mass diagrams (CSMDs) at four different redshifts and the stellar mass dependence of the colour cuts we applied to divide the red and blue galaxies. In each panel, the colour map indicates the distribution of the logarithmic comoving number density of galaxies in cells of $(g - r)^{0.1}$ and M_*, normalized by the stellar mass function (SMF) at that M_*. This normalization highlights the $g - r$ ranges with relatively high concentration of galaxies along the colour axis, enhancing the appearance of the 'red sequence' on each panel. The CSMDs are cut off at different stellar masses due to the redshift dependence of the mixture limit, which is ultimately related to the flux limit of the spectroscopic survey. The red dashed lines going through the red sequences are uniform across all redshifts, indicating that the loci of the red sequence on the CSMD has little redshift dependence within our sample. To divide the galaxies into red and blue, we therefore define the colour cut to be parallel to the red sequence on the CSMD, described by

$$(g - r)^{0.1}_{\rm cut}(M_*) = 0.8\left(\frac{\lg M_*}{10.5}\right)^{0.6}, \quad (2)$$

and indicated by the black solid lines in Fig. 1. The weak stellar mass dependence in equation (2) causes a variation of $(g - r)^{0.1}_{\rm cut}$ between 0.76 and 0.84 within our sample, leading to differences in classification between blue versus red of only a few per cent compared to a constant cut at 0.8.

[1] http://home.strw.leidenuniv.nl/∼jarle/SDSS/

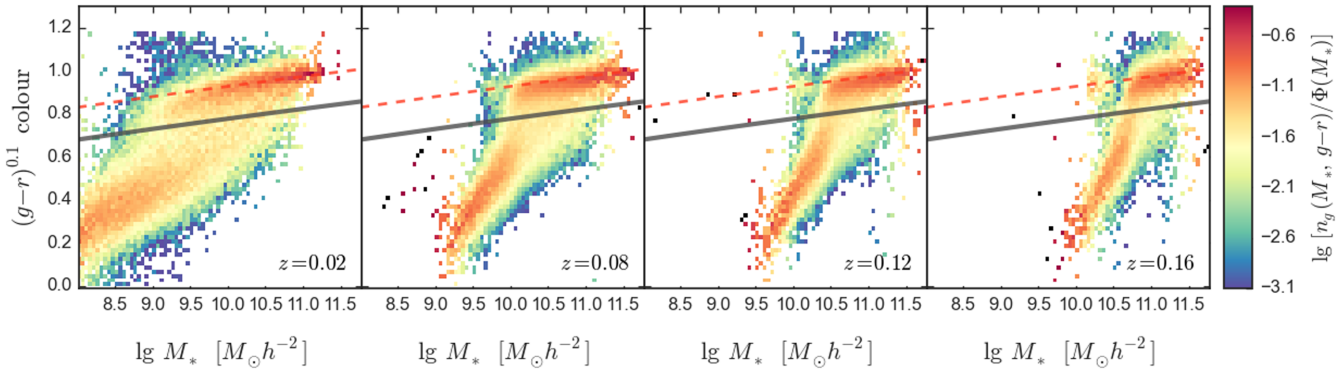

Figure 1. The colour–stellar mass diagram of the SDSS MGS at four different redshifts (from left to right: $z = 0.02$, 0.08, 0.12, and 0.16). The colour maps indicate the comoving galaxy number density at fixed colour and M_*, normalized by the SMF at that M_*. The red dashed lines are the same across all panels, indicating little redshift evolution in the locus of red sequence on the colour–M_* diagram. The black solid lines are the redshift-independent colour cut that we use to divides the galaxies into red and blue populations.

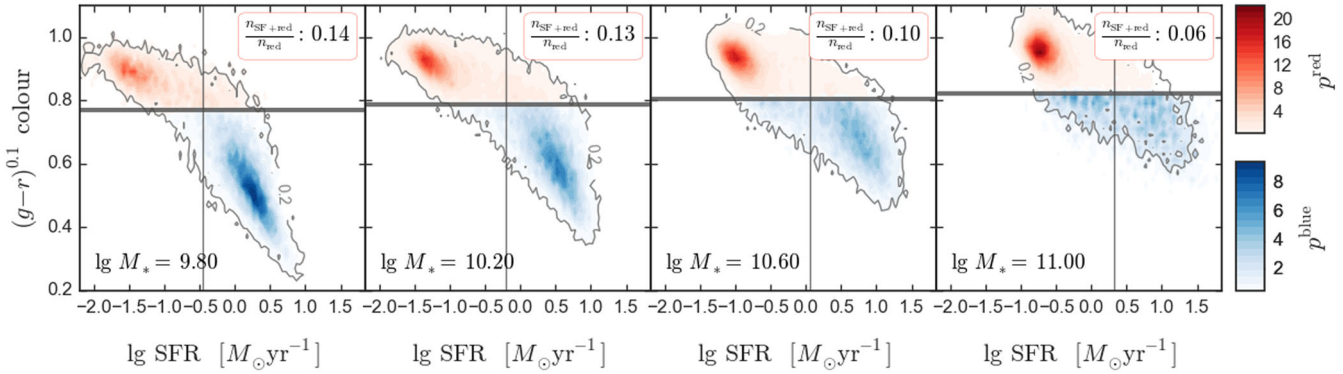

Figure 2. Illustration of the agreement between using $(g - r)^{0.1}$ colour and SFR as quenching indicators. Each panel shows the joint PDFs of the colour and the total SFR of red and blue galaxies at four different stellar masses (from left to right: lg $M_* = 9.8$, 10.2, 10.6, and 11.0). Grey horizontal thick and vertical thin lines indicate the colour and the SFR cuts that divide the galaxies into red/blue and passive/active, respectively. The PDFs of red and blue galaxies, each normalized to unity, are represented by two separate sets of contour maps with their levels indicated by the corresponding colour bars on the right. The red galaxy samples include a very small fraction of dusty star-forming galaxies, with a contamination rate of ∼6–14 per cent depending on stellar mass (marked on the top right of each panel).

Whether a galaxy is quiescent or star forming, however, is never a clear-cut choice. Galaxy bimodality shows in nearly every aspect of galaxy properties, including broad-band colour, star formation rate (SFR), morphology (e.g. late/early-type, De Vaucouleurs/exponential profile), and concentration (e.g. Sérsic index). Bernardi et al. (2010) found that many late-type (Sb and later) galaxies lie above the red galaxy colour cut (similar to equation 2) and they tend to be edge-on discs reddened by dust. Conversely, some early-type galaxies lie below the cut, either showing star-forming AGN or post-starburst spectrum, with their star formation history well described by a recent minor and short starburst superimposed on old stellar component (Huang & Gu 2009). As a result, Woo et al. (2013) advocated the use of SFR as the quenching indicator, and they claimed that one third of the red galaxies are star forming. However, the large fraction of star-forming contaminants in the 'red' population in Woo et al. (2013) is mainly caused by the rest-frame $U - B$ colour Woo et al. (2013) adopted in defining red galaxies, derived from *AB* magnitudes that are *K*-corrected from the SDSS *ugriz* photometry. Using the native $(g - r)^{0.1}$ colour largely eliminates the star formers from the red galaxies.

Fig. 2 clearly demonstrates the good consistency between using $(g - r)^{0.1}$ colour and SFR as quenching indicators. The four panels show the joint 2D probability density distributions (PDFs) of galaxy colour and the logarithmic SFR at four different M_* (from left to right: lg $M_* = 9.8$, 10.2, 10.6, and 11.0). In each panel, the thick horizontal line represents the colour cut defined in equation (2), while the thin vertical line indicates the SFR value that saddles the separate SFR distributions of passive and active galaxies, which can be well described by

$$\lg \frac{{\rm SFR}_{\rm cut}}{M_*} = -0.35(\lg M_* - 10.0) - 10.23, \quad (3)$$

in parallel to the star-forming sequence defined in Salim et al. (2007). The fraction of dusty, star-forming galaxies in the red population decreases from 0.14 to 0.06 as stellar mass increases from lg $M_* = 9.8$ to 11.0, significantly lower than the one third reported in Woo et al. (2013) using $U - B$ colours. Assuming that the dusty, star-forming galaxies reside in similar haloes as the regular blue, star-forming ones, their contamination of the 'red' samples would dilute the intrinsic difference in the clustering and lensing between the quiescent and active galaxies by only several per cent. Therefore, we conclude that it is robust to use red fraction as a proxy for quenching efficiency, and the results of our analysis should stay the same if SFR were used for the selection of quenched galaxies.

As described in Paper I, the `iHOD` model constructs individual HODs within very narrow redshift slices (we use $\Delta z = 0.01$), so that

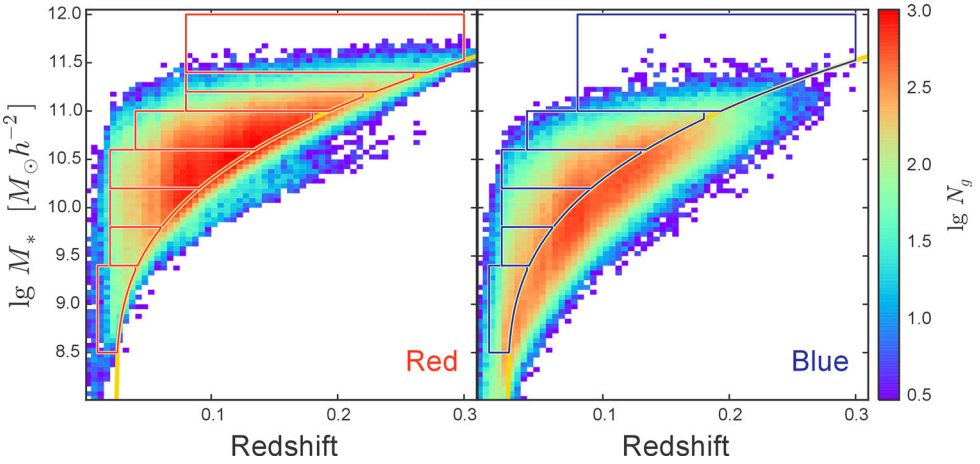

Figure 3. Selection of the red (left) and the blue (right) galaxy samples on the M_*–redshift diagram. In each panel, the colour map indicates the logarithmic number counts of galaxies at fixed M_* and redshift. The thick yellow curve is the 'mixture limit' above which the completeness of the galaxy samples is relatively high. By selecting almost all galaxies above this mixture limit, our iHOD analysis is able to include 84 per cent more than galaxies than the traditional HOD methods.

Table 1. Red and blue stellar mass bins used for the iHOD quenching analysis, corresponding to the selections in the left- and right-hand panels of Fig. 3, respectively. The red selection includes three stellar mass samples above lg $M_* = 11$ as opposed to the single lg $M_* > 11$ sample in the blue selection, while the five lower stellar mass red and blue samples share the same binning in lg M_* and z.

lg($M_*/h^{-2}\,\mathrm{M}_\odot$)	z	N^{red}	N^{blue}
8.5–9.4	0.01–0.04	3224	10 773
9.4–9.8	0.02–0.06	7336	9356
9.8–10.2	0.02–0.09	28 301	19 883
10.2–10.6	0.02–0.13	70 514	29 160
10.6–11.0	0.04–0.18	84 108	21 058
11.0–11.2	0.08–0.22	22 626	
11.2–11.4	0.08–0.26	9775	
11.4–12.0	0.08–0.30	2875	
11.0–12.0	0.08–0.30		4095

the sample selection does not require a single uniform stellar mass range among all the redshift slices within that sample, i.e. having a rectangular shape on the M_*–z diagrams. Fig. 3 illustrates the galaxy samples selected on the M_*–z diagram within each coloured population for the iHOD quenching analysis. The colour intensity represents the logarithmic galaxy number counts in cells of M_* and z. As mentioned in Section 2.1, all selected samples have the 'wedge'-like stellar mass thresholds described by the mixture limit, and thus contain extra galaxies at the far end of the redshift range that are usually unused in traditional HOD analysis. Additionally, since those high-redshift wedges have a larger comoving volume per unit redshift than the low redshifts, they include the most abundant regions on the M_*–z diagram, corresponding to the reddest regions on both panels of Fig. 3. The resultant increase in the selected galaxy sample sizes is more than 80 per cent compared to traditional selections.

Above the mixture limit, the red galaxies (left-hand panel) are two times more abundant than the blue galaxies (right-hand panel), despite the fact that the ratio of the two colours is close to unity in the spectroscopic survey. Therefore, we can afford finer binning in M_* in the red galaxy sample than in the blue galaxy sample, especially at the high-mass end. Table 1 summarizes the basic information of the two sets of sample selections used by the iHOD quenching analysis. In total, we divide 228 759 red galaxies and 94 325 blue galaxies into eight and six subsamples, respectively.

2.3 Measuring galaxy clustering w_p and g–g lensing $\Delta\Sigma$

We measure the projected correlation function w_p for each galaxy sample by integrating the 2D redshift–space correlation function ξ^s,

$$w_\mathrm{p}(r_\mathrm{p}) = \int_{-r_\pi^{\max}}^{+r_\pi^{\max}} \xi^s(r_\mathrm{p}, r_\pi)\mathrm{d}r_\pi\,, \tag{4}$$

where r_p and r_π are the projected and the line-of-sight (LOS) comoving distances between two galaxies. We measure the w_p signal out to a maximum projected distance of $r_\mathrm{p}^{\max} = 20\,h^{-1}$ Mpc, where the galaxy bias is approximately linear. For the integration limit, we adopt a maximum LOS distance of $r_\pi^{\max} = 60\,h^{-1}$ Mpc. We only use the w_p values down to the physical distance that corresponds to the fibre radius at the maximum redshift of each sample, with fewer data points for higher stellar mass (hence larger maximum fibre radius) samples.

The Landy–Szalay estimator (Landy & Szalay 1993) is employed for computing the 2D correlation $\xi^s(r_\mathrm{p}, r_\pi)$. The error covariance matrix for each w_p measurement is estimated via the jackknife resampling technique. We divide the entire footprint into 200 spatially contiguous, roughly equal-size patches on the sky and compute the w_p for each of the 200 jackknife subsamples by leaving out one patch at a time. For each stellar mass sample, we adopt the sample mean of the 200 subsample measurements as our final estimate of w_p, and the sample covariance matrix as an approximate to the underlying error covariance.

For the surface density contrast $\Delta\Sigma$, we measure the projected mass density in each radial bin by summing over lens–source pairs 'ls' and random lens–source pairs 'rs',

$$\Delta\Sigma(r_\mathrm{p}) = \langle\Sigma_\mathrm{crit}\gamma_\mathrm{t}(r_\mathrm{p})\rangle = \frac{\sum_\mathrm{ls} w_\mathrm{ls} e_\mathrm{t}^{(\mathrm{ls})}\Sigma_\mathrm{crit}(z_\mathrm{l}, z_\mathrm{s})}{2\mathcal{R}\sum_\mathrm{rs} w_\mathrm{rs}}, \tag{5}$$

where e_t is the tangential ellipticity component of the source galaxy with respect to the lens position, the factor of $2\mathcal{R}$ converts our definition of ellipticity to the tangential shear γ_t, and w_ls is the inverse variance weight assigned to each lens–source pair (including

shot noise and measurement error terms in the variance). $\Sigma_{\rm crit}$ is the so-called critical surface mass density, defined as

$$\Sigma_{\rm crit}^{-1}(z_{\rm l}, z_{\rm s}) \equiv \frac{4\pi G}{c^2} \frac{D_{\rm ls} D_{\rm l} (1+z_{\rm l})^2}{D_{\rm s}}, \quad (6)$$

where $D_{\rm l}$ and $D_{\rm s}$ are the angular diameter distances to lens and source, and $D_{\rm ls}$ is the distance between them. We use the estimated photometric redshift each source to compute $D_{\rm s}$ and $D_{\rm ls}$. The factor of $(1+z_{\rm l})^2$ comes from our use of comoving coordinates. We subtract off a similar signal measured around random lenses, to subtract off any coherent systematic shear contributions (Mandelbaum et al. 2005); this signal is statistically consistent with zero for all scales used in this work. Finally, we correct a bias in the signal caused by the uncertainties in the photometric redshift using the method from Nakajima et al. (2012).

To calculate the error bars, we also used the jackknife re-sampling method. As shown in Mandelbaum et al. (2005), internal estimators of error bars (in that case, bootstrap rather than jackknife) perform consistently with external estimators of error bars for $\Delta\Sigma$ on small scales due to its being dominated by shape noise.

3 QUENCHING MODELS AND SIGNAL PREDICTIONS

In this section, we introduce the mathematical descriptions of the hybrid (Section 3.1) and the halo (Section 3.2) quenching models. We also briefly describe how to infer the `iHOD`s for the red and blue galaxies in Section 3.3, but refer reader to Paper I for more details on the `iHOD` framework. The prediction of $w_{\rm p}$ and $\Delta\Sigma$ from each coloured `iHOD` is rather complex but exactly the same as that for the overall galaxy populations, therefore we directly refer readers to the relevant sections (4 and 5) in Paper I for details. We ignore quenching via mergers in both quenching models considered below, as merging-induced quenching is negligible at $z < 0.5$ (P10).

3.1 Hybrid quenching model

The hybrid quenching model parametrizes the red fraction as a function of both M_* and $M_{\rm h}$, aiming to mimic the empirical stellar mass and environment quenching trends observed in P10. In the physical picture implied by this model, every quiescent galaxy had spent some portion of its life on the star-forming 'main sequencing' as a central before the eventual quenching (Daddi et al. 2007; Noeske et al. 2007; Speagle et al. 2014), due to either the depletion of gas supply (i.e. stellar mass quenching) or entering another halo as a satellite, when environment quenching kicked in.

While the stellar mass trend is straightforward to parametrize, it is unclear whether the environment quenching trend among satellites can be mimicked by a trend in halo mass, as the environment–halo mass relation is very complex and depends strongly on the definition of that environment. The P10 environment of satellites, as defined by $\delta_{\rm 5NN}$, shows strong correlation with group richness when the richness is below five. In richer systems, however, the correlation is mostly smeared out and $\delta_{\rm 5NN}$ instead anticorrelates with the halo-centric distance $D_{\rm g}$ (Peng et al. 2012). This apparent transition between the two richness regimes is caused by the increase of $D_5/R_{\rm vir}$, the ratio between the typical distance to the fifth nearest neighbour to the halo virial radius, from below to above unity. When $D_5 > R_{\rm vir}$, $\delta_{\rm 5NN}$ is roughly proportional to $M_{\rm h}/D_5^3$, thus more tied to $M_{\rm h}$. When $D_5 < R_{\rm vir}$, $\delta_{\rm 5NN}$ is essentially an intra-halo overdensity measured at $D_{\rm g}$, which depends more strongly on $D_{\rm g}$ than $M_{\rm h}$ due to the steep declining slope of the Navarro–Frenk–White (NFW)-like halo density profile.

But for our purposes, what matters is the *mean* satellite quenching efficiency as a function of halo mass $M_{\rm h}$, averaged over galaxies at all $D_{\rm g}$ within that halo. As pointed out by Woo et al. (2013), the density profile of more massive haloes falls off less steeply with distance than that of less massive systems, so the probability of finding the fifth nearest neighbour increases with halo mass. Therefore, the P10 environment quenching trend can be potentially encapsulated within the halo model as a satellite quenching dependence on halo mass.

Assuming stellar mass as the main driver of central galaxy quenching, we parametrize the red fraction of centrals as

$$f_{\rm cen}^{\rm red}(M_*, M_{\rm h}) \equiv 1 - g(M_*) = 1 - \exp\left[-\left(M_*/M_*^{\rm q}\right)^{\mu}\right], \quad (7)$$

where $M_*^{\rm q}$ is a characteristic stellar mass ($f_{\rm cen}^{\rm red}(M_*^{\rm q}, M_{\rm h}) = (e-1)/e = 0.632$) and μ dictates how fast the quenching efficiency increases with M_*, with $\mu = 1$ being exponential. The satellites are subject to an extra halo quenching term $h(M_{\rm h})$, so that

$$f_{\rm sat}^{\rm red}(M_*, M_{\rm h}) = 1 - g(M_*)h(M_{\rm h}), \quad (8)$$

with

$$h(M_{\rm h}) = \exp\left[-\left(M_{\rm h}/M_{\rm h}^{\rm q}\right)^{\nu}\right], \quad (9)$$

where $M_{\rm h}^{\rm q}$ is a characteristic halo mass and ν controls the pace of satellite quenching. The above equations, including g and h as powered exponential functions, are very similar to the fitting formula adopted in Baldry et al. (2006) and Peng et al. (2012).

The top left and right panels of Fig. 4 illustrate the central and satellite red fractions, computed from the best-fitting hybrid quenching model via equations (7) and (8), respectively. The arrow in each panel points in the direction of increasing quenching efficiency on the 2D plane of M_* and $M_{\rm h}$. For the central galaxies, although the quenching is driven by M_* along the horizontal axis, the red fraction still shows strong increasing trend with $M_{\rm h}$ due to the tight correlation between M_* and $M_{\rm h}$, i.e. the SHMR of central galaxies. The 2D distribution of satellite red fractions displays a 'boxy' pattern, echoing the separate stellar mass and environment quenching trends detected in P10.

Combining the central and the satellite terms, the red fraction of galaxies with stellar mass M_* inside haloes of total mass $M_{\rm h}$ is

$$\begin{aligned} f^{\rm red}(M_*, M_{\rm h}) &= f_{\rm sat}(M_*, M_{\rm h})\, f_{\rm sat}^{\rm red}(M_*, M_{\rm h}) \\ &\quad + [1 - f_{\rm sat}(M_*, M_{\rm h})]\, f_{\rm cen}^{\rm red}(M_*, M_{\rm h}), \end{aligned} \quad (10)$$

where $f_{\rm sat}(M_*, M_{\rm h})$ is the satellite fraction that can be predicted by the overall `iHOD` model from Paper I. For the hybrid model, equation (10) can be reduced to

$$f^{\rm red}(M_*, M_{\rm h}) = g(M_*)\{1 - f_{\rm sat}(M_*, M_{\rm h})[1 - h(M_{\rm h})]\}. \quad (11)$$

3.2 The halo quenching model

As described in the Introduction, the halo quenching model relies on halo mass alone to quench both central and satellite galaxies, and Gabor & Davé (2015) demonstrated that it also naturally explains the stellar mass and environment quenching trends seen in P10, by embedding galaxies in massive haloes filled with hot gas via virial heating. However, depending on the exact physical processes driven by $M_{\rm h}$, halo quenching may apply to the central and satellites differently. For instance, while the central galaxies in haloes above $M_{\rm shock}$ could be quenched by shocked-heated gas and then maintain a

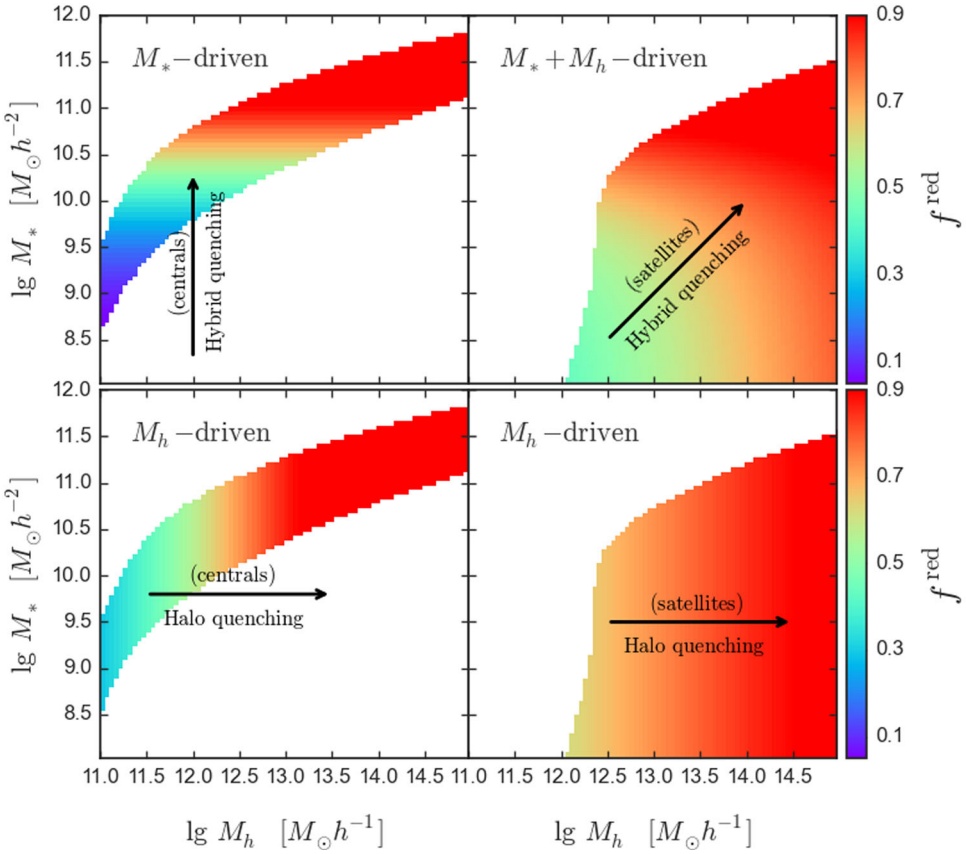

Figure 4. Illustration of the difference between hybrid (top row) and halo (bottom row) quenching models. For each quenching model, we show the red galaxy fraction as a function of both stellar mass and halo mass for the central (left) and satellite (right) populations, respectively, predicted from the best-fitting parameters. The long arrow in each panel indicates the direction of increasing red fraction, driven by either stellar mass (top left), halo mass (bottom left and right), or the combination of both (top right) within the two quenching models.

high gas temperature via the 'raido'-mode feedback from AGNs, the satellite galaxies in the those haloes may still retain some cold gas as the 'central' galaxies of its own coherent sub-group. Therefore, the satellite galaxies continue to accrete gas and convert it to stars over a period of ∼1 Gyr after entering into a larger halo (Simha et al. 2009). Similar processes like slow strangulation (assuming no accretion on to satellites) also produce prolonged quenching actions (Peng, Maiolino & Cochrane 2015). In this case, the halo quenching of centrals and satellites are somewhat decoupled, and the quenching of satellites is a more gradual process than that of centrals.

Therefore, unlike the hybrid model, we describe the red fractions of centrals and satellites as two independent functions of $M_{\rm h}$:

$$f_{\rm cen}^{\rm red}(M_*, M_{\rm h}) = 1 - f_{\rm cen}^{\rm blue}(M_{\rm h}) = 1 - \exp\left[-\left(M_{\rm h}/M_{\rm h}^{\rm qc}\right)^{\mu^{\rm c}}\right], \quad (12)$$

and

$$f_{\rm sat}^{\rm red}(M_*, M_{\rm h}) = 1 - f_{\rm sat}^{\rm blue}(M_{\rm h}) = 1 - \exp\left[-\left(M_{\rm h}/M_{\rm h}^{\rm qs}\right)^{\mu^{\rm s}}\right], \quad (13)$$

where $M_{\rm h}^{\rm qc}$ and $M_{\rm h}^{\rm qs}$ are the critical halo masses responsible for triggering quenching of central and satellites, respectively, and $\mu^{\rm c}$ and $\mu^{\rm s}$ are the respective powered-exponential indices controlling the transitional behaviour of halo quenching across the critical halo masses.

Similarly, the bottom left and right panels of Fig. 4 illustrate the central and satellite red fractions, computed from the best-fitting halo quenching model via equations (12) and (13), respectively, with arrows indicating halo mass as the sole driver for quenching in both populations. The orthogonality of the hybrid and halo quenching directions for central galaxies is the key distinction that we look to exploit in this paper, by identifying its imprint on the clustering and g–g lensing signals of red and blue galaxies. The total red fraction can be obtained by substituting equations (12) and (13) into equation (10).

Finally, we emphasize that in reality the true quenching arrow could be pointing anywhere between the two orthogonal directions, i.e. a more generalized quenching model consisting of a linear mixture of the two, with the linear coefficients varying as functions of M_* and $M_{\rm h}$ as well – schematically,

$$Q_{\rm true} = \omega(M_*, M_{\rm h}) \times Q_{\rm hybrid} + (1 - \omega(M_*, M_{\rm h})) \times Q_{\rm halo}. \quad (14)$$

However, as a first step of constraining quenching, the goal of this paper is to find out if ω is closer to zero (i.e. halo-quenching dominated) or unity (i.e. hybrid-quenching dominated).

3.3 From quenching models to colour-segregated `iHOD`s

In order to predict the $w_{\rm p}$ and $\Delta\Sigma$ for the red and blue galaxies in each quenching model, we construct `iHOD`s for both coloured populations by combining the overall `iHOD` with $f^{\rm red}(M_*, M_{\rm h})$ predicted by that quenching model.

Let us start with the red galaxies. The key is to derive $p^{\rm red}(M_*, M_{\rm h})$, the 2D joint PDF of the red galaxies of stellar mass M_* sitting in haloes of mass $M_{\rm h}$, given the 2D PDF of the overall galaxy population $p(M_*, M_{\rm h})$ inferred from Paper I,

$$p^{\rm red}(M_*, M_{\rm h}) = \frac{f^{\rm red}(M_*, M_{\rm h})}{f^{\rm red}_{\rm tot}}\, p(M_*, M_{\rm h}), \tag{15}$$

where $f^{\rm red}_{\rm tot}$ is the overall red faction of all galaxies, obtained via

$$f^{\rm red}_{\rm tot} = \iint f^{\rm red}(M_*, M_{\rm h}) p(M_*, M_{\rm h})\, {\rm d}M_{\rm h}{\rm d}M_*. \tag{16}$$

As described in Paper I, iHOD predicts the $w_{\rm p}$ and $\Delta\Sigma$ signals for a given galaxy sample by combining the predicted signals from individual narrow redshift slices, each of which is described by a single standard HOD.

For deriving standard HODs within redshift slices for the red galaxies, we need

$$p^{\rm red}(M_{\rm h}|M_*) = \frac{p^{\rm red}(M_{\rm h}, M_*)}{p^{\rm red}(M_*)}, \tag{17}$$

while $p^{\rm red}(M_*)$ is the predicted *parent* (i.e. including observed and unobserved galaxies) SMF of the red galaxies normalized by their total number density $n^{\rm red}_{\rm tot}$,

$$p^{\rm red}(M_*) = \frac{\phi^{\rm red}(M_*)}{n^{\rm red}_{\rm tot}} = \int_0^{+\infty} p^{\rm red}(M_{\rm h}, M_*)\, {\rm d}M_{\rm h}. \tag{18}$$

Finally, we arrive at the HOD of red galaxies at any redshift z as

$$\left\langle N^{\rm red}(M_{\rm h}|z)\right\rangle = \left(\frac{{\rm d}n}{{\rm d}M_{\rm h}}\right)^{-1} \int_{M_*^0}^{M_*^1} p^{\rm red}(M_{\rm h}|M_*)\, \Phi^{\rm red}_{\rm obs}(M_*|z)\, {\rm d}M_*, \tag{19}$$

where $\Phi^{\rm red}_{\rm obs}(M_*|z)$ is the *observed* SMF of red galaxies at redshift z, directly accessible from the survey. For modelling the samples defined in Fig. 3 for the iHOD analysis, we measure the observed galaxy SMF at each redshift, and then obtain the HOD for that redshift slice using equation (19). In this way, we avoid the need to explicitly model the sample incompleteness as a function of M_* and/or $M_{\rm h}$.

For the blue galaxies, we apply the same procedures above to obtain $\langle N^{\rm blue}(M_{\rm h}|z)\rangle$ from $p^{\rm blue}(M_{\rm h}, M_*)$, by substituting $f^{\rm red}(M_*, M_{\rm h})$ with $f^{\rm blue}(M_*, M_{\rm h})\equiv 1 - f^{\rm red}(M_*, M_{\rm h})$ in equations (15)–(19).

Fig. 5 illustrates the two sets of coloured iHODs derived from the best-fitting hybrid (top row) and halo (bottom row) quenching models. In each row, the left- and right-hand panels display $\lg({\rm d}N(M_*|M_{\rm h})/{\rm d}\lg M_*)$, the average log number of galaxies per dex in stellar mass within haloes at fixed mass, for the red and blue populations, respectively. The white and black contour lines highlight the central and satellite galaxy occupations separately on the M_*–$M_{\rm h}$ plane. All panels reveal the same generic pattern, consisting of a tight sequence that corresponds to the SHMR of the central galaxies, and a cloud underneath occupied by the satellite galaxies. The level of similarity exhibited by the red galaxies is especially high between the two quenching models (left-hand column).

However, comparing the left- and right-hand panels in the same row (i.e. red versus blue galaxies in the same quenching model), the red centrals are more preferentially sitting in the high-M_* and high-$M_{\rm h}$ region than in the low-M_* and low-$M_{\rm h}$ one, while the opposite is true for the blue centrals. This segregation happens regardless of quenching models, confirming our notion that it is difficult to unambiguously disentangle the two quenching directions, despite their orthogonality, by merely examining the quenching trend with M_*, or some surrogate of $M_{\rm h}$ that has substantial scatter about the true $M_{\rm h}$ (e.g. group richness).

The satellites are quenched by $M_{\rm h}$ in both models, but are also partially by M_* in the hybrid model. Thus, there are more high-M_* blue satellite galaxies within massive haloes in the halo quenching model (bottom-right panel) than in the hybrid model (top-right panel). In addition, the low-mass haloes in the hybrid quenching model are more likely to host blue dwarf satellites than in the halo quenching model.

The two sets of iHOD models, presented in this section and in Fig. 5, are the analytical foundation that allow us to predict the $w_{\rm p}$ and $\Delta\Sigma$ signals as functions of the four parameters in each quenching model. Any difference shown in Fig. 5 between the two models will be propagated to the different behaviours in the final predictions of $w_{\rm p}$ and $\Delta\Sigma$, and is thus detectable by comparing the two sets of predictions to the measurements from SDSS galaxies.

4 CONSTRAINTS ON THE TWO QUENCHING MODELS

4.1 Constraints of the quenching parameters

Ideally one would constrain both the iHOD parameters and the quenching parameters together, by simultaneously fitting to the $w_{\rm p}$ and $\Delta\Sigma$ measurements of the overall, red, and blue galaxies. However, since the measurements of the overall population have the highest signal-to-noise ratio and the overall iHOD does not include quenching, it is conceptually more reasonable to adopt a two-step scheme. In the first step, we constrain the iHOD parameters using only the measurements of the galaxy samples without dividing by colour (i.e. Paper I). In the second step, when constraining the quenching parameters, we either fix the best-fitting iHOD parameters (i.e. the 'fixed case') or input the iHOD constraints from Paper I as priors (i.e. the 'prior case'). In particular, for the prior case we draw the iHOD parameters from the joint prior distribution represented by the Markov Chain Monte Carlo samples derived in Paper I. For each quenching model, we adopt the results from the prior case as our fiducial constraint in the following analysis.

In addition to the powerful statistical features of the iHOD framework inherited from Paper I, our quenching analysis also adds two important advantages compared to the traditional HOD modelling of red and blue galaxies. First, traditional HOD studies of red and blue galaxies treat the two populations independently, so that the total number of HOD parameters inevitably doubles compared to the modelling of the overall galaxy population (e.g. Tinker et al. 2013; Rodríguez-Puebla et al. 2015). In our analysis, the red and blue populations are derived not from scratch, but by filtering the overall iHOD with the red fraction predicted by each quenching model, which is described by only four simple yet physically meaningful parameters. Our method also guarantees that the sum of the red and blue SHMRs is mathematically identical to the overall SHMR. Secondly, the traditional method usually parametrizes the red galaxy fraction as a 1D function of halo mass, while our method affords a 2D function of $f^{\rm red}$ defined on the M_*–$M_{\rm h}$ plane, which is crucial to the task of examining stellar mass as a potential driver for quenching.

For each quenching model, we infer the posterior probability distributions of the four model parameters from the $w_{\rm p}$ and $\Delta\Sigma$ measurements of the eight red and six blue galaxy samples within a Bayesian framework. We model the combinatorial vector $\boldsymbol{x}$ of the $w_{\rm p}$ and the $\Delta\Sigma$ components of the red and blue galaxies as a

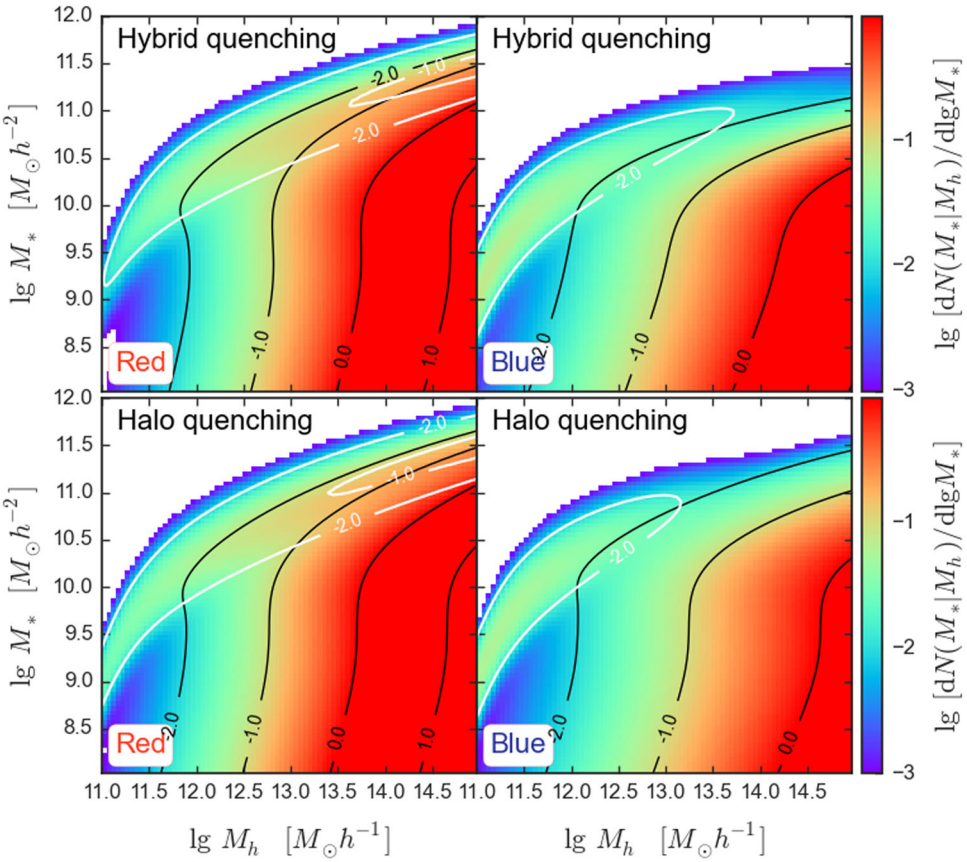

Figure 5. The iHOD, i.e. the average number of galaxies per dex in stellar mass within haloes at fixed mass of the red (left-hand column) and blue (right-hand column) populations, predicted from the best-fitting hybrid (top row) and halo (bottom row) quenching models based on the same iHOD of the overall population. White and black contour lines in each panel indicate the iHOD distributions for the central and satellite galaxies, respectively, while the colour contours show the iHOD for the total galaxy population. The two best-fitting models produce similar overall patterns of galaxy occupation, with very subtle but important differences – compared to hybrid quenching, halo quenching produces stronger segregation between red and blue centrals in high versus low-mass haloes.

multivariate Gaussian, which is fully specified by its mean vector ($\bar{\boldsymbol{x}}$) and covariance matrix ($\mathbf{C}$). The Gaussian likelihood is thus

$$\mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta}) = |\mathbf{C}|^{-1/2} \exp\left(-\frac{(\boldsymbol{x}-\bar{\boldsymbol{x}})^T \mathbf{C}^{-1} (\boldsymbol{x}-\bar{\boldsymbol{x}})}{2}\right), \tag{20}$$

where

$$\boldsymbol{\theta} \equiv \{\lg M_*^{\rm q}, \mu, \lg M_{\rm h}^{\rm q}, \nu\} \tag{21}$$

in hybrid quenching, and

$$\boldsymbol{\theta} \equiv \{\lg M_{\rm h}^{\rm qc}, \mu^{\rm c}, \lg M_{\rm h}^{\rm qs}, \mu^{\rm s}\} \tag{22}$$

in halo quenching.

We adopt flat priors on the model parameters, with a uniform distribution over a broad interval that covers the entire possible range of each parameter (see the second column of Table 2). The final covariance matrix $\mathbf{C}$ is assembled by aligning the error matrices of $w_{\rm p}$ and $\Delta\Sigma$ measured for individual coloured samples along the diagonal blocks of the full $N \times N$ matrix. We ignore the weak covariance between $w_{\rm p}$ and $\Delta\Sigma$ (with the covariance being weak due to the fact that $\Delta\Sigma$ is dominated by shape noise), and between any two measurements of the same type but for different stellar mass or coloured samples.

Fig. 6 presents a summary of the inferences from the halo quenching model analysis, showing the 1D posterior distribution for each of the four model parameters (diagonal panels), and the 95 and 68 per cent confidence regions for all the parameter pairs (off-diagonal panels). In the panels of the lower triangle, we highlight the results from our fiducial model, i.e. the prior case, employing the iHOD parameter constraints from Paper I as priors. In each panel of the upper triangle, we compare the constraints from the fiducial analysis (filled contours) to that of the fixed case analysis, which keeps the iHOD parameters at their best-fitting values derived from Paper I. The two analyses are consistent with each other, implying that the explanation of the red and blue signals does not require any modification in the description of the overall galaxy population. The two inferred characteristic halo mass scales are very similar to the critical shock heating mass scale, $M_{\rm h}^{\rm qc} \sim M_{\rm h}^{\rm qs} \sim M_{\rm shock}$, while the two powered-exponential indices, $\mu^{\rm c}$ and $\mu^{\rm s}$, indicate that the central and the satellite quenching transition differently across that shared characteristic halo mass. We defer the detailed discussion of the physical implications of the halo quenching constraints in Section 6. The 68 per cent confidence regions of the 1D posterior constraints are listed in Table 2.

Similarly, Fig. 7 presents the constraints on the hybrid quenching model. The critical stellar mass for quenching all galaxies is $M_*^{\rm q} = 3.16(\pm 0.31) \times 10^{10}\, h^{-2}\, {\rm M}_\odot$, echoing the characteristic stellar mass for downsizing at the low redshift. The stellar mass quenching index μ is slightly below unity, the value required for maintaining the observed redshift-independence of Schechter M^* and faint-end slope of the star-forming galaxies in the stellar mass

Table 2. Description, prior specifications, and posterior constraints of the parameters in the halo (top) and hybrid (bottom) quenching models. All the priors are uniform distributions running across the entire range of possible values for the parameters, and the uncertainties are the 68 per cent confidence regions derived from the 1D posterior probability distributions.

Parameter	Description	Uniform prior range	Prior case	Fixed case
	Halo quenching model $Q_{\rm halo}$			
$\lg M_{\rm h}^{\rm qc}$ [$h^{-1}\,{\rm M}_\odot$]	Characteristic halo mass for central galaxy quenching.	[11.0, 15.5]	$12.20^{+0.07}_{-0.08}$	$12.25^{+0.05}_{-0.06}$
$\mu^{\rm c}$	Pace of central galaxy quenching with halo mass.	[0.0, 3.0]	$0.38^{+0.04}_{-0.03}$	$0.42^{+0.03}_{-0.03}$
$\lg M_{\rm h}^{\rm qs}$ [$h^{-1}\,{\rm M}_\odot$]	Characteristic halo mass for satellite galaxy quenching.	[11.0, 15.5]	$12.17^{+0.12}_{-0.10}$	$12.30^{+0.17}_{-0.23}$
$\mu^{\rm s}$	Pace of satellite galaxy quenching with halo mass.	[0.0, 3.0]	$0.15^{+0.03}_{-0.02}$	$0.16^{+0.03}_{-0.03}$
	Hybrid quenching model $Q_{\rm hybrid}$			
$\lg M_*^{\rm q}$ [$h^{-2}\,{\rm M}_\odot$]	Characteristic stellar mass for central and satellite quenching.	[9.0, 12.0]	$10.50^{+0.04}_{-0.04}$	$10.55^{+0.03}_{-0.03}$
μ	Pace of galaxy quenching with stellar mass.	[0.0, 3.0]	$0.69^{+0.06}_{-0.06}$	$0.66^{+0.06}_{-0.05}$
$\lg M_{\rm h}^{\rm q}$ [$h^{-1}\,{\rm M}_\odot$]	Characteristic halo mass for satellite galaxy quenching.	[11.0, 15.5]	$13.76^{+0.15}_{-0.14}$	$13.63^{+0.10}_{-0.11}$
ν	Pace of satellite galaxy quenching with halo mass.	[0.0, 3.0]	$0.15^{+0.05}_{-0.05}$	$0.18^{+0.05}_{-0.04}$

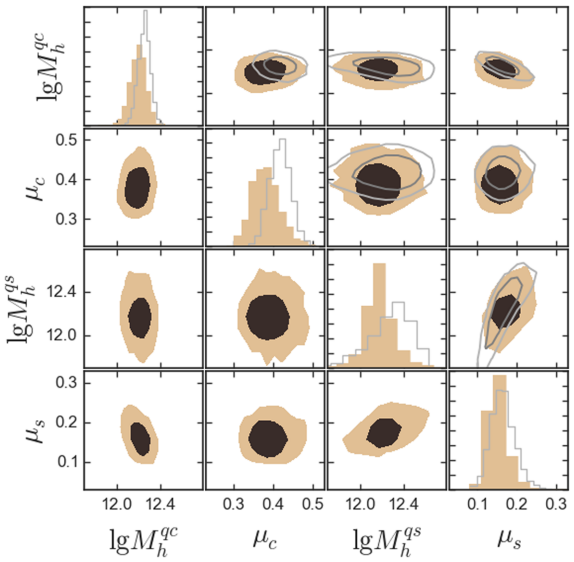

Figure 6. Confidence regions from our halo quenching model analysis of the clustering and lensing of the red and blue galaxies in the 2D planes that comprised of all the pair sets of the four quenching parameters. Histograms in the diagonal panels show 1D posterior distributions of individual parameters. Contour levels run through confidence limits of 95 per cent (light brown) and 68 per cent (dark brown) inwards. The filled contours show the constraints from our fiducial model where the `iHOD` parameters that describe the overall galaxy population are drawn from priors informed by the analysis in Paper I, while the open contours in the panels of the upper triangle are the constraints from a simpler model where the overall `iHOD` parameters are kept fixed at the best-fitting values derived from Paper I.

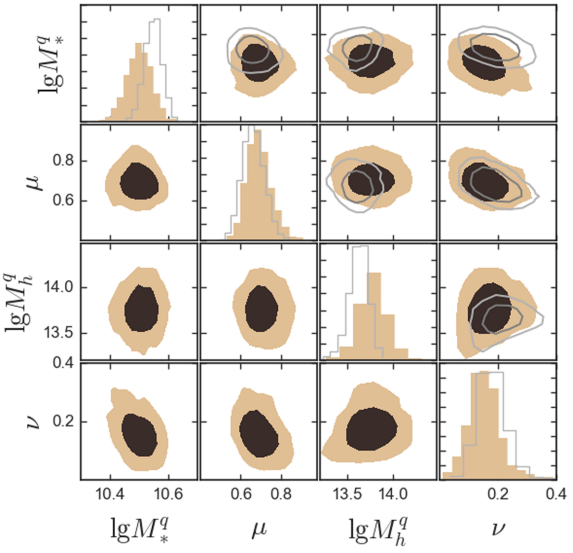

Figure 7. Similar to Fig. 6, but for the hybrid quenching model.

quenching formalism proposed in P10. The characteristic halo mass for the quenching of satellites is much higher than $M_{\rm h}^{\rm qs}$, albeit with a similar quenching index of $\nu = 0.15 \pm 0.05$.

4.2 Best-fitting model predictions

Fig. 8 compares the clustering (top row) and g–g lensing (bottom row) signals measured from SDSS (points with error bars) to those predicted by the best-fitting halo (solid lines) and hybrid (dashed line) quenching models, for the eight red (left-hand column) and the six blue (right-hand column) stellar mass samples. In terms of the overall goodness-of-fit, the best-fitting halo quenching model yields a χ^2 of 701.0, while the hybrid quenching model has a worse χ^2 of 736.8. The reduced χ^2 values are thus 1.60 and 1.68 for the halo and hybrid quenching, respectively, both providing reasonable fits to the data, considering that the uncertainties in the measurements of the low-M_* samples are underestimated. We defer a discussion of the statistical significance of both best fits to the upcoming section.

For the red galaxy samples, the two quenching models predict very similar signals except for the two lowest stellar mass bins. Unfortunately the $w_{\rm p}$ measurements in these two bins are severely affected by the underestimated cosmic variance due to the small volumes, with highly correlated uncertainties on all scales. Therefore, neither quenching model gives an adequate fit to their $w_{\rm p}$ signals. The $\Delta\Sigma$ signals of the two lowest mass bins are less affected by cosmic variance because the measurement error is dominated by shape noise, and are thus better described by the two quenching model predictions.

The two quenching models also predict very similar signals for the blue galaxies, except for the high-mass ones with $\lg M_*{>}11$. While both quenching models give adequate fits to the $w_{\rm p}$ signals of these massive blue galaxies, the halo quenching model produces much better fit to their $\Delta\Sigma$ signals than the hybrid quenching model, driving most of the difference in the log-likelihoods (i.e. the χ^2 values) of the two best-fitting models. This difference revealed by

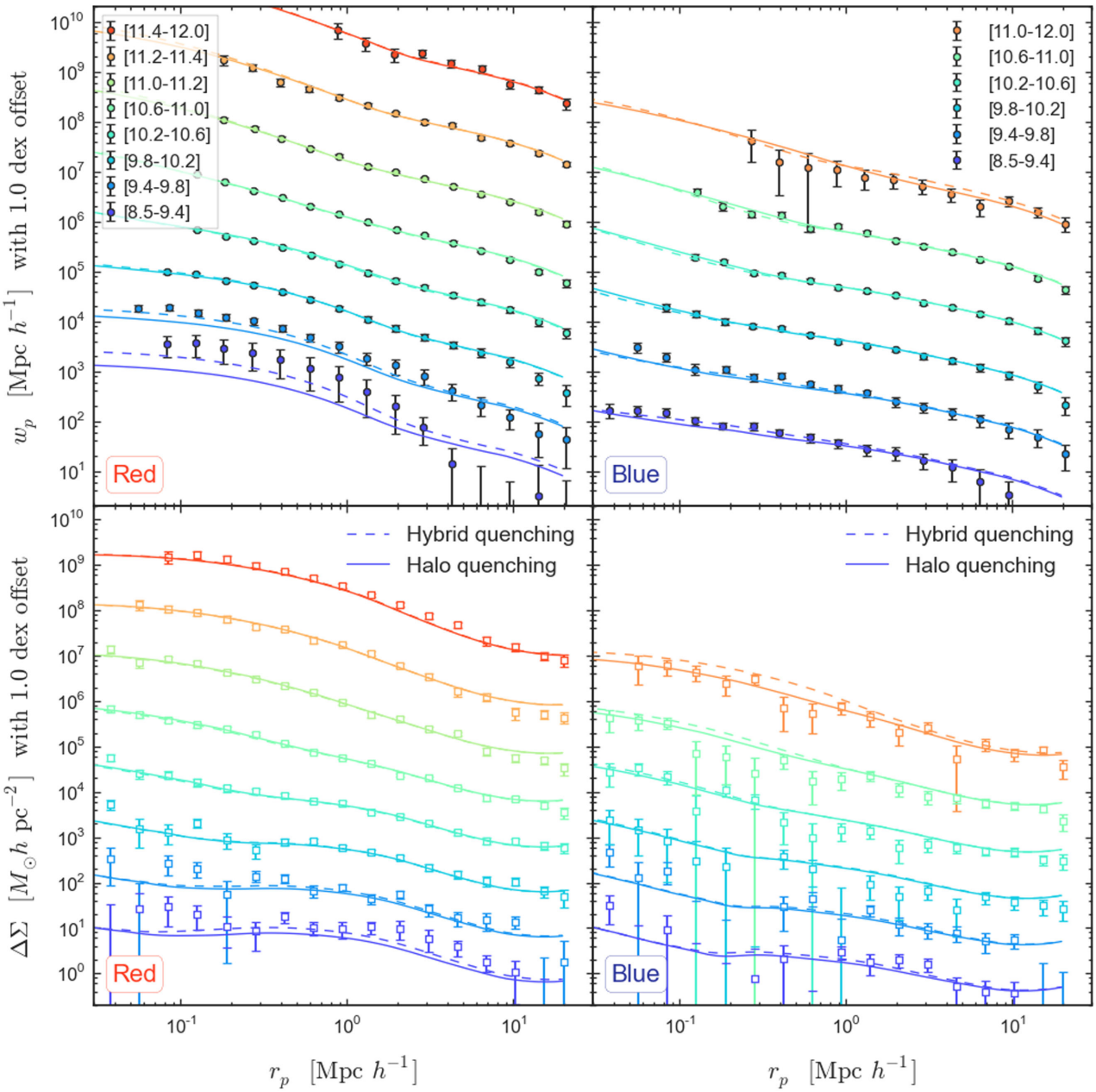

Figure 8. Comparison between the measurements of clustering (top row) and lensing (bottom row) signals to the predictions by the two best-fitting quenching models, for the eight red stellar mass samples (left-hand column) and the six blue samples (right-hand column). In each panel, the signals from bottom to top are in ascending order of the average stellar mass of the galaxy samples, each shifted from its adjacent sample by 1.0 dex to avoid clutter. Points with error bars are the measurements, while the solid and the dashed curves are the predictions from the best-fitting halo and hybrid quenching models, respectively. The two best-fitting models provide equally adequate fits to the measurements for the red galaxies, but predict different g–g lensing signals of high-M_* blue galaxies (bottom right; the two high-mass samples).

the massive blue galaxies, as will be discussed further later, is the key to distinguishing the two quenching models.

Fig. 9 highlights the split between the red and blue galaxies from the overall population in the w_p (left) and $\Delta\Sigma$ (right) signals, predicted by the two best-fitting quenching models for the eight stellar mass bins marked in the right-hand panel. In each panel, the thick grey curves are the `iHOD` predictions for the overall galaxy samples, which bifurcate into the thin red and blue curves, i.e. predictions for the red and blue galaxies. Solid and dashed line styles indicate the halo and hybrid quenching models, respectively. As seen in Fig. 8, the two quenching models predict very similar bifurcation signatures, except for the high-mass bins where the hybrid quenching predicts a stronger large-scale bias, a weaker small-scale clustering strength, but a stronger small-scale g–g lensing amplitude, than the halo quenching for the blue galaxies. Unfortunately the measured w_p signals for the high-mass galaxies are cut off at small scales due to fibre collision, and the measurement uncertainties in the large-scale w_p are not small enough to distinguish the two quenching predictions (top-right panel of Fig. 8).

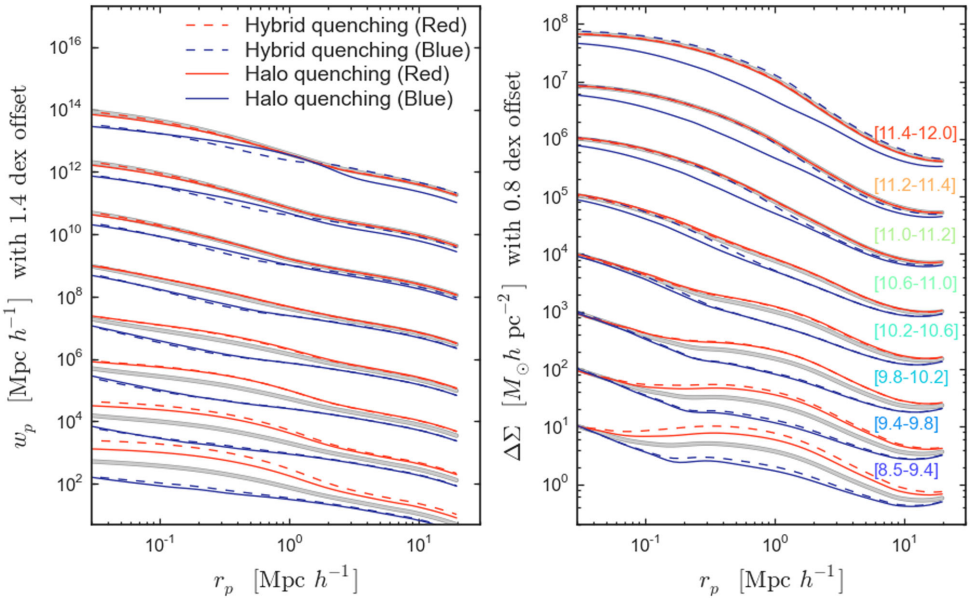

Figure 9. Comparison between the clustering (left) and lensing (right) signals predicted by the halo (solid) and hybrid (dashed) quenching models for the red (red), and blue (blue) galaxies. The thick grey curves indicate the best-fitting `iHOD` model prediction for the overall galaxy population, with the stellar mass ranges marked on the right-hand panel. The two quenching models generally predict similar scale and stellar mass dependences of the red and blue deviations from predictions of the overall population (comparing solid and dashed lines of the same colour), except for the high-mass blue lenses (comparing the blue solid lines to the blue dashed lines in the right-hand panel).

Therefore, the g–g lensing of the massive blue galaxies clearly provides the most discriminative information, as shown in the right-hand panel of Fig. 9. For blue galaxies above $10^{11}\,h^{-2}\,\mathrm{M}_\odot$, the halo quenching model predicts substantially lower weak lensing amplitudes than the hybrid model on all distance scales, and thus provides a much better fit to the measurements (see bottom-right panel of Fig. 8).

To understand the discrepancy between the two quenching predictions for the massive blue galaxies, we show the decomposition of w_p (top row) and $\Delta\Sigma$ (bottom row) signals predicted by the best-fitting hybrid (left-hand column) and halo (right-hand column) quenching models for the lg $M_* = 11.0$–12.0 blue sample in Fig. 10. In each panel, the orange data points with error bars and the thick blue curve are the measured and predicted signals for the blue sample, while the `iHOD` prediction for the overall sample is shown by the thin green curve. The best-fitting quenching model prediction is then decomposed into contributions from the 1-halo and 2-halo (thin dotted) terms. For w_p the 1-halo term includes the contributions from central–satellite pairs (thin solid; '1-h c–s') and satellite–satellite (thin dashed; '1-h s–s') pairs; For $\Delta\Sigma$ the 1-halo term consists of a satellite term (thin dashed) and a non-satellite term (thin solid). We also include a point source stellar mass term in $\Delta\Sigma$, which is model-independent and negligible on most of the relevant scales (not shown here). Most importantly, the 1-halo non-satellite term is directly related to the average dark matter density profile of the host haloes (including both the main haloes for centrals and the subhaloes for satellites), and its amplitude is proportional to the average mass of those haloes. The halo quenching model clearly provides a much better fit to the data than the hybrid model, with factors of 2 and 4 improvement in χ^2 for w_p and $\Delta\Sigma$, respectively. In addition, Fig. 10 shows the crucial advantage of including g–g lensing in the joint analysis – since the best-fitting hybrid quenching model adequately describes the galaxy clustering ($\chi^2/N = 1.07$), its deficiency would not be exposed unless we compare the $\Delta\Sigma$ predictions to data ($\chi^2/N = 3.84$).

Fig. 10 reveals two major differences between the two quenching model predictions: (1) the halo quenching model predicts a much higher satellite fraction among the massive blue galaxies than the hybrid model, hence the more prominent '1-halo satellite' terms; and (2) the average (sub)halo mass of those massive blue galaxies predicted by the halo quenching model is much lower compared to the hybrid model prediction, hence the lower g–g lensing amplitudes and better fit to the data. Roughly speaking, since the hybrid quenching model relies on the stellar mass to quench central galaxies, it tends to place central galaxies at fixed M_* into similar haloes regardless of their colours. However, in the halo quenching model any galaxies that are unquenched have to live in lower mass haloes than their quenched counterparts with similar M_*. In the section below we will argue that, the discrepancy between the average halo masses of the massive blue galaxies predicted by the two quenching models is insensitive to the details in the model parameters, therefore can be used as a robust feature for identifying the dominant quenching driver.

4.3 Origin of host halo mass segregation between red and blue centrals

Comparison between the two best-fitting predictions in Section 4.2 reveals that $\langle M_\mathrm{h}|M_*\rangle$, the average host halo mass at fixed stellar mass (i.e. the mean halo-to-stellar mass relation; HSMR), is potentially the key discriminator of the two types of quenching models. In particular, by predicting a lower $\langle M_\mathrm{h}|M_*\rangle$ for the blue centrals, the halo quenching model provides a much better fit to the w_p and $\Delta\Sigma$ signals of the massive blue galaxies than the hybrid quenching model. But before going any further, we need to understand the cause of this discrepancy between the two quenching models,

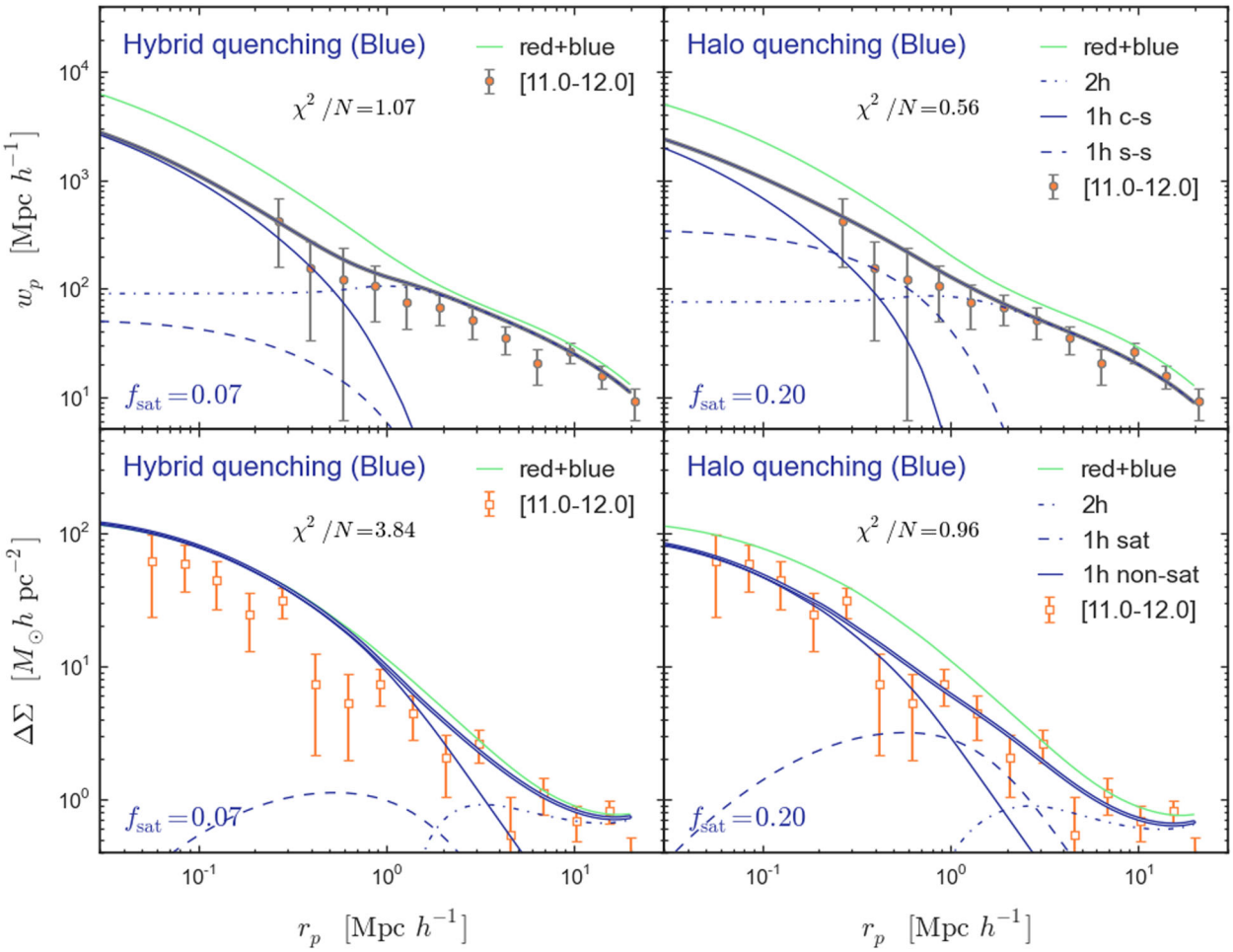

Figure 10. Decomposition of the clustering (top) and lensing (bottom) signals of the high-mass blue galaxy sample (lg $M_* = [11.0, 12.0]$), by the best-fitting hybrid (left) and halo (right) quenching models. In each panel, the green curve shows the predicted signal for the overall galaxy population at this stellar mass. The measurement and the predicted signal for blue galaxies are shown by the orange points with error bars and the thick blue curve, respectively, with the reduced χ^2 marked on the top. The predicted signal is then decomposed into different components involving separate contributions from centrals and satellites (see text for details). In particular, the '1-halo non-satellite' terms in the $\Delta\Sigma$ panels indicate the underlying dark matter density profiles of the host (sub)haloes. The halo quenching model provides a much better fit to the $w_{\rm p}$ and $\Delta\Sigma$ measurements by allowing substantially fewer massive haloes to host these high-M_* blue galaxies as centrals, and a higher satellite fraction among these massive blue galaxies.

especially to answer the following questions. First, what is the origin of the host halo mass segregation between the two colours? Secondly, is the halo quenching necessary for predicting the strong segregation in $\langle M_{\rm h}|M_*\rangle$ between the red and blue centrals, and can the stellar mass quenching process produce an equally low halo mass for the massive blue centrals with a different μ?

For red or blue central galaxies, the conversion from the mean SHMR (i.e. $\langle M_*|M_{\rm h}\rangle$) to its inverse relation, the HSMR, is highly non-trivial. Using the blue centrals as an example, the HSMR can be computed from

$$\langle M_{\rm h}|M_*\rangle^{\rm blue}_{\rm cen} = \int p^{\rm blue}_{\rm cen}(M_{\rm h}|M_*) M_{\rm h}\, {\rm d}M_{\rm h}, \tag{23}$$

where

$$p^{\rm blue}_{\rm cen}(M_{\rm h}|M_*) = \frac{p^{\rm blue}_{\rm cen}(M_*|M_{\rm h}) p^{\rm blue}_{\rm cen}(M_{\rm h})}{p^{\rm blue}_{\rm cen}(M_*)} \propto p^{\rm blue}_{\rm cen}(M_*|M_{\rm h}) f^{\rm blue}_{\rm cen}(M_*, M_{\rm h}) \frac{{\rm d}n(M_{\rm h})}{{\rm d}M_{\rm h}}. \tag{24}$$

In the above equation, $p^{\rm blue}_{\rm cen}(M_*|M_{\rm h})$ is the PDF of blue central galaxy stellar mass at fixed $M_{\rm h}$, determined by the mean SHMR of the blue centrals and its scatter, $f^{\rm blue}_{\rm cen}$ is the blue fraction of centrals, and ${\rm d}n/{\rm d}M_{\rm h}$ is the halo mass function. Therefore, for given cosmology the HSMR of the blue central galaxies has two components, the blue central SHMR (both mean and scatter) and the blue fraction of centrals. To understand $\langle M_{\rm h}|M_*\rangle$ for both colours more quantitatively, we start by examining the red and blue SHMRs predicted by the two models.

The top and bottom panels in the left-hand column of Fig. 11 show the mean SHMRs of the total, red, and blue central galaxies, predicted by the best-fitting hybrid and halo quenching models, respectively. Coloured bands indicate the logarithmic scatters about the mean relations. The hybrid quenching model predicts a segregation in M_* between the red and blue central galaxies at fixed halo mass, as the high M_* galaxies are more likely to be quenched. The halo quenching model, however, predicts exactly the same SHMRs for all three populations, as galaxies at fixed halo mass are equally likely to be quenched regardless of stellar mass. The red and blue segregation in M_*, or the lack thereof, is best illustrated in the two

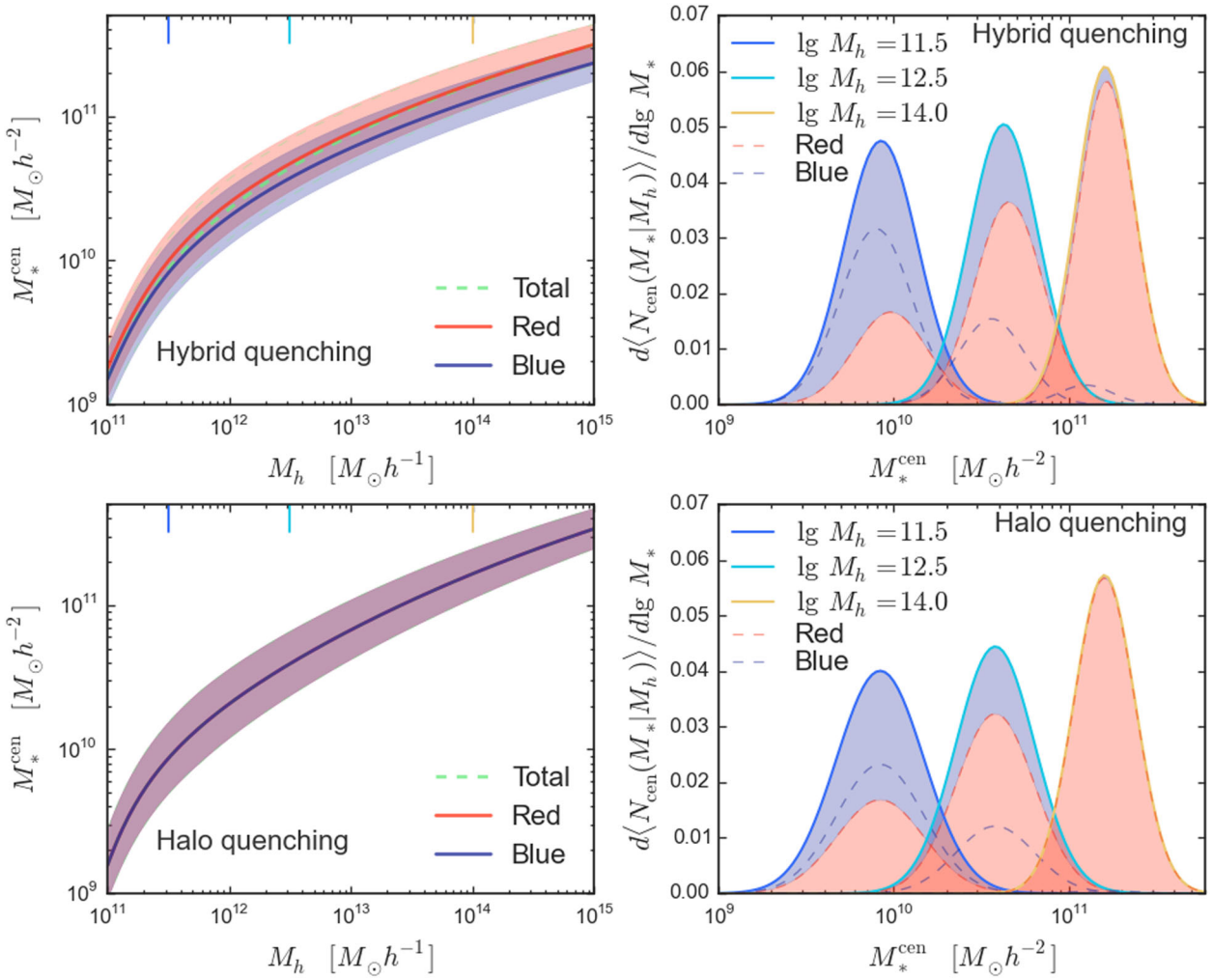

Figure 11. Comparison between the SHMRs derived by the hybrid (top) and halo quenching (bottom) models. On the left-hand panel of each row, red solid, blue solid, and green dashed curves indicate the average logarithmic stellar mass of red, blue, and any central galaxies at fixed halo mass, respectively, predicted by the corresponding quenching model. The shaded band of each colour shows the lognormal scatter about the average relation. The three coloured ticks on the top indicates the three halo masses for which we show the central galaxy stellar mass distributions in the right-hand panel. The integral of the distribution for each halo mass (shown in the legend) is the expected total number of central galaxies at that halo mass, which can be decomposed into the contributions from red (red shaded histogram) and blue (blue dashed histogram or blue shaded area) centrals. The halo quenching model produces exactly the same SHMRs for the two coloured populations as the overall relation, due to the lack of correlation between quenching and stellar mass at fixed halo mass, whereas the hybrid quenching model makes fractionally more red centrals at higher M_*.

right-hand panels of Fig. 11, using three halo masses as examples (lg $M_* = 11.5, 12.5, 14.0$).

In each panel, the total filled area for each halo shows the stellar mass distribution of central galaxies in that halo. The width of the distribution decreases with halo mass due to the flattening of SHMR on the high-mass end. Under each distribution, the red and blue shaded areas represent the contributions from the red and blue centrals, so that the sum of the red and blue SHMRs exactly recovers the total SHMR. In the hybrid quenching model for any given halo mass, the red galaxy distributions are shifted to higher M_* compared to the blue distributions, which are indicated by the dashed histograms and are equivalent to the blue shaded regions. The halo quenching model produces zero such shift. The non-zero shift in the hybrid model drives the SHMR of the blue central galaxies to become shallower than that of the red centrals as seen in the top-left panel of Fig. 11. Naively, one might think that this shift will also cause the blue centrals to reside in more massive haloes than the red centrals if we simply compare the inverse functions of the two SHMRs – a shallower (blue) SHMR maps the same M_* on the y-axis to a higher halo mass on the x-axis.

However, a more careful inspection of the segregation patterns reveals a second, and much more important difference in the predicted fraction of blue galaxies among centrals – blue centrals persist in all halo masses in the hybrid quenching model, but barely show up in the $10^{14}\,h^{-1}\,{\rm M}_\odot$ haloes in the halo quenching model. The left-hand panel of Fig. 12 illustrates the blue fractions as functions of $M_{\rm h}$ predicted by the best-fitting halo (thick black solid) and hybrid (thick blue dashed) quenching models. The amplitude of $f_{\rm cen}^{\rm blue}$ in hybrid quenching also depends on M_* and the blue dashed curve is the average blue fraction over all galaxies above $10^8\,h^{-2}\,{\rm M}_\odot$. While $f_{\rm cen}^{\rm blue}$ in the halo quenching case strictly follows the powered exponential form (i.e. equation 12), in the hybrid case it is affected by both the stellar mass quenching and the slope of the SHMR. We also show the blue fraction of satellites (thin solid) derived by the halo quenching analysis, which exhibits a slower decline with $M_{\rm h}$ compared to that of centrals.

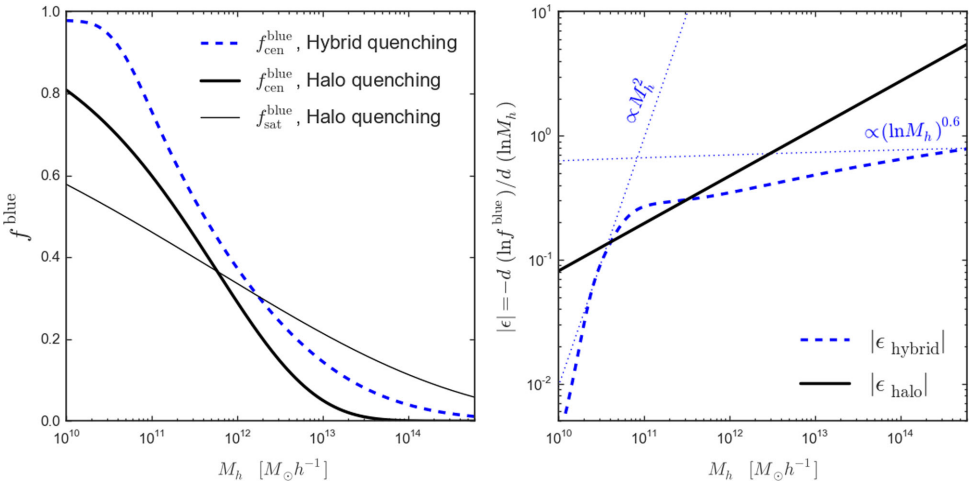

Figure 12. Left-hand panel: blue galaxy fractions as functions of halo mass predicted by the two best-fitting quenching models. Black and blue thick solid curves are the blue fractions within centrals $f_{\rm cen}^{\rm blue}(M_{\rm h})$ predicted by the best-fitting halo and hybrid quenching models, respectively. Black thin curve indicates the blue satellite fraction predicted by the best-fitting halo quenching model. Right-hand panel: the magnitude of the logarithmic slopes of $f_{\rm cen}^{\rm blue}(M_{\rm h})$, characterized by $|\epsilon(M_{\rm h})|$ in the two quenching models. The hybrid quenching model predicts a rapid decline of $f_{\rm cen}^{\rm blue}$ at low masses and then slows drastically at high masses, while the halo quenching model maintains a steady decline over all masses. The two dotted lines are the asymptotic behaviours of the hybrid model at very low and high masses, derived in equation (31). The slow decline of $f_{\rm cen}^{\rm blue}$ in the hybrid model overpredicts the fraction of blue centrals among massive haloes.

The blue galaxy fractions of centrals decline rapidly with halo mass in both quenching models, but the speed of decline varies differently as a function of halo mass between the two models. To investigate this quantitatively, we define the 'central galaxy quenching rate' as a function of halo mass, $\epsilon(M_{\rm h})$, as the logarithmic rate at which the blue fraction declines with halo mass, $\mathrm{d}\ln f_{\rm cen}^{\rm blue}/\mathrm{d}\ln M_{\rm h}$, which is shown in the right-hand panel of Fig. 12 for each model. As expected, the halo quenching produces a steady increase of $|\epsilon|$ with $M_{\rm h}$

$$\epsilon_{\rm halo} \equiv \left(\frac{\mathrm{d}\ln f_{\rm cen}^{\rm blue}}{\mathrm{d}\ln M_{\rm h}}\right)_{\rm halo} \propto -M_{\rm h}^{\mu^c} \simeq -M_{\rm h}^{0.35}. \tag{25}$$

For the hybrid quenching case, $f_{\rm cen}^{\rm blue}(M_{\rm h})$ experiences a rapid decline at low $M_{\rm h}$ and then a gradual one at high $M_{\rm h}$. This shift in gear can be understood as follows. The central galaxy quenching rate $\epsilon_{\rm hybrid}$ depends on both the $f_{\rm cen}^{\rm blue}(M_*)$ and the derivative of the SHMR, so that

$$\epsilon_{\rm hybrid} \equiv \left(\frac{\mathrm{d}\ln f_{\rm cen}^{\rm blue}}{\mathrm{d}\ln M_{\rm h}}\right)_{\rm hybrid} = \left(\frac{\mathrm{d}\ln f_{\rm cen}^{\rm blue}}{\mathrm{d}\ln M_*}\right)_{\rm hybrid} \left(\frac{\mathrm{d}\ln M_*}{\mathrm{d}\ln M_{\rm h}}\right). \tag{26}$$

Since $f_{\rm cen}^{\rm blue}(M_*)$ also has a powered exponential form (see equation 7),

$$\left(\frac{\mathrm{d}\ln f_{\rm cen}^{\rm blue}}{\mathrm{d}\ln M_*}\right)_{\rm hybrid} \propto -M_*^{\mu}. \tag{27}$$

The slope of the SHMR $\mathrm{d}\ln M_*/\mathrm{d}\ln M_{\rm h}$ is tightly constrained by Paper I, which described the SHMR $f_{\rm SHMR} \equiv \exp\langle \ln M_*(M_{\rm h})\rangle$ as the inverse of

$$\ln\frac{M_{\rm h}}{M_1} = \beta\ln m + \left(\frac{m^\delta}{1+m^{-\gamma}} - \frac{1}{2}\right), \tag{28}$$

where $m \equiv M_*/M_{*,0}$, $M_{*m,0}{\sim}2\times10^{10}\,h^{-2}\,\mathrm{M}_\odot$ and $M_1 \sim 1.3\times10^{12}\,h^{-1}\,\mathrm{M}_\odot$ are the characteristic stellar and halo mass that separate the behaviours in the low- and high-mass ends, and the remaining parameters control the running slopes of the SHMR. Assuming reasonable values of the slope parameters (i.e. $\beta\sim0.33$, $\delta\sim0.42$, $\gamma\sim1.21$; see Paper I), equation (28) can be approximated by

$$\ln\frac{M_{\rm h}}{M_1} \simeq \begin{cases} \beta\ln m - 0.5, & m \ll 1 \\ m^\delta + [\beta\ln m - 0.5], & m \gg 1. \end{cases} \tag{29}$$

Clearly, the SHMR is a steep power-law relation at the low-mass end, with $M_* \propto M_{\rm h}^{1/\beta} \sim M_{\rm h}^3$, whereas at the high-mass end the slope of SHMR is very shallow, with $M_* \propto (\ln M_{\rm h})^{1/\delta} \sim (\ln M_{\rm h})^{2.4}$.

Therefore, the slope of the SHMR is

$$\frac{\mathrm{d}\ln M_*}{\mathrm{d}\ln M_{\rm h}} \simeq \begin{cases} 1/\beta, & M_{\rm h} \ll M_1 \\ (\delta\ln M_{\rm h})^{-1}, & M_{\rm h} \gg M_1. \end{cases} \tag{30}$$

Combining equations (26), (27) and (30), we arrive at

$$\epsilon_{\rm hybrid} \propto \begin{cases} -M_{\rm h}^{\mu/\beta}, & M_{\rm h} \ll M_1 \\ -(\ln M_{\rm h})^{(\mu-\delta)/\delta}, & M_{\rm h} \gg M_1. \end{cases} \tag{31}$$

Assuming $\mu = 0.67$ from the best-fitting hybrid quenching model, we have

$$\epsilon_{\rm hybrid} \propto \begin{cases} -M_{\rm h}^2, & M_{\rm h} \ll M_1 \\ -(\ln M_{\rm h})^{0.6}, & M_{\rm h} \gg M_1. \end{cases} \tag{32}$$

The above equation is shown as the dotted lines on the right-hand panel of Fig. 12, roughly reproducing the two distinctive asymptotic behaviours of $\epsilon_{\rm hybrid}$ at the low- and high-mass ends. The actual slope of $\epsilon_{\rm hybrid}$ is steeper than predicted by equation (32) at high masses, where equation (30) becomes less accurate.

The comparison between $\epsilon_{\rm halo}$ and $\epsilon_{\rm hybrid}$ in Fig. 12 (i.e. equations 32 and 25) clearly reveals that, the halo quenching model does not quench central galaxies in the low-mass haloes as efficiently as the hybrid model, but by maintaining a steady quenching rate at $\epsilon(M_{\rm h}) \sim -M_{\rm h}^{0.35}$ the halo quenching model is able to quench almost all centrals in the very massive haloes. The hybrid quenching model, however, is relatively inefficient to quench massive central galaxies in the very high mass haloes. When calculating the HSMR using equation (24), this difference in ϵ completely dominates the effect due to the slight difference between the two coloured SHMRs. Therefore, the stellar mass quenching, due to its slow central galaxy

quenching rate on the high-mass end, is incapable of producing a strong segregation in the HSMR between the two colours. In order for the hybrid quenching model to mimic the steeper slope of $\epsilon_{\rm halo}(M_{\rm h})$, the stellar mass quenching trend would have to drop so precipitously that the abundance of blue galaxies is cut off beyond some maximum stellar mass, which is ruled out by the observed SMFs of blue galaxies (see Fig. 14). Therefore, we further emphasize that this slow quenching rate with halo mass in the hybrid model is caused by the changing slope of SHMR across M_1, and is thus insensitive to the stellar mass quenching prescriptions, e.g. the value of μ.

To summarize the findings above using the quenching diagram of Fig. 4, the steep slope ($M_*\sim M_{\rm h}^3$) of the SHMR below M_1 makes the SHMR more aligned with the quenching arrow along the M_*-axis (i.e. stellar mass quenching, see top-left panel of Fig. 4), causing progressively more galaxies to be quenched at higher halo mass. Above M_1, however, the SHMR becomes shallower ($M_*\sim(\ln M_{\rm h})^{2.4}$) and is almost perpendicular to the quenching arrow, leaving a substantial number of blue centrals in massive haloes. As a result, the massive blue centrals are extremely scarce in the $10^{14}\,h^{-1}\,{\rm M}_\odot$ haloes in the halo quenching model (bottom-right panel of Fig. 11), but have a much stronger presence within those haloes in the hybrid quenching model (top-right panel of Fig. 11). By the same token, a strong segregation in the host halo mass between the red and blue centrals would points to the necessity of a dominant halo mass quenching for the central galaxies.

5 COMPARING THE HYBRID AND HALO QUENCHING MODELS

In this section, we perform a robust comparison between the two quenching models in two ways, an internal one based on Bayesian Information Criterion (BIC) described in Section 5.1, and an external one based on cross-validation (Section 5.2), which is motivated by the quenching impact on the average halo mass of the massive blue galaxies quantitatively explained in Section 4.3.

5.1 Internal model comparison: BIC

In Bayesian applications, pairwise comparisons between models M_1 and M_2 are often based on the Bayes factor B_{12}, which is defined as the ratio of the posterior odds, $P(M_1|\boldsymbol{x})/P(M_2|\boldsymbol{x})$, to the prior odds, $\pi(M_1)/\pi(M_2)$. In our case, the Bayes factor is

$$B_{12}=\frac{P(Q_{\rm halo}|\boldsymbol{x})}{P(Q_{\rm hybrid}|\boldsymbol{x})}\frac{\pi(Q_{\rm hybrid})}{\pi(Q_{\rm halo})}, \tag{33}$$

so that a B_{12} above unity indicates the data favour halo quenching and a B_{12} below points to hybrid quenching. However, in most practical settings (as is the case here) the prior odds are hard to set precisely, and model selection based on BIC is widely employed as a rough equivalent to selection based on Bayes factors. The BIC (aka, 'Schwarz information criterion'), is defined as

$${\rm BIC}=-2\ln\mathcal{L}_{\rm max}+k\ln n, \tag{34}$$

where $\ln\mathcal{L}_{\rm max}$ is the maximum likelihood value, k is the number of parameters, and n is the number of data points. Kass & Raftery (1995) argued that in the limit of large n ($n=439$ in our analyses),

$$\frac{-2\ln B_{12}-({\rm BIC}_{\rm halo}-{\rm BIC}_{\rm hybrid})}{-2\ln B_{12}}\longrightarrow 0, \tag{35}$$

i.e. $\Delta{\rm BIC}={\rm BIC}_{\rm halo}-{\rm BIC}_{\rm hybrid}$ can be viewed as a rough approximation to $-2\ln B_{12}$, so that $\Delta{\rm BIC}<0(<-10)$ indicates that $Q_{\rm halo}$ is favoured (strongly) and $\Delta{\rm BIC}>0(>10)$ points (strongly) to $Q_{\rm hybrid}$.

The ΔBIC between the two quenching models is -35.8, which corresponds to an asymptotic value of $\ln B_{12}=17.9$ according to equation (35). Therefore, based on the BIC test, the clustering and the g–g lensing measurements of the red and blue galaxies strongly favour the halo quenching model against the hybrid quenching model, and the halo mass is the more statistically dominant driver of galaxy quenching than stellar mass.

The two quenching models are non-nested models with the same k and n, so the second term of equation (34) that penalizes model complexities is the same in both quenching models. The BIC test is then equivalent to the alternative Akaike information criterion that is based on relative likelihoods, or a simple $\Delta\chi^2$ test (i.e. $\Delta\chi^2=35.8$). These tests all point to the halo mass as the main driver of quenching.

5.2 External model comparison: halo masses of massive blue centrals

The discussion in Section 4.3 points to a potentially smoking-gun test of the two quenching models, by comparing the host halo mass of the massive red and blue central galaxies predicted from the two best-fitting quenching models, to other mass measurements for observed groups/clusters with red and blue centrals within the same redshift range. Unfortunately, clusters with blue centrals are systematically underselected by most of the photometric cluster finders based on matching to the red sequence, while spectroscopic group catalogues constructed from friends-of-friends algorithms do not have large enough volume for finding many massive clusters.

Recently, Mandelbaum et al. (2015) constructed a sample of locally brightest galaxies (LBGs) from the SDSS MGS, by adopting a set of isolation criteria carefully calibrated against semi-analytic mock galaxy catalogues to minimize the satellite contamination rate (Wang et al. 2016). The resulting LBG sample is thus a subset of all massive central galaxies, but with excellent purity of central galaxy membership and zero bias against blue colour. Therefore, the LBGs are ideal for our purpose of measuring the segregation in halo mass between the red and blue centrals.

Mandelbaum et al. (2015) measured the average host halo mass of the LBGs directly by fitting an NFW density profile (after projection to 2D) to the weak lensing signals measured below $1\,h^{-1}$ Mpc. Fig. 13 compares the host halo mass measured as a function of LBG stellar mass (data points with error bars) to that predicted by the best-fitting halo (solid) and hybrid (dashed) quenching models. The error bars on the LBG measurements are the 1σ uncertainties on the mean halo mass, derived from 1000 bootstrap-resampled data sets. The coloured bands about the solid curves are the uncertainties on the mean halo mass predicted from the 68 per cent confidence regions. To avoid clutter, we do not show the uncertainties on the hybrid quenching predictions, which are comparable to the halo quenching uncertainties. The average halo mass predicted by the halo quenching model is in excellent agreement with the measurements from the LBG sample, while the hybrid quenching model, as expected, grossly overpredicts the halo mass for the massive blue galaxies.

The difference in the host halo mass between red and blue centrals, as predicted by the hybrid quenching model, can be understood simply as the outcome of a larger differential growth between dark matter and stellar mass in quiescent systems than in star-forming ones. More specifically, in quenched systems the dark matter haloes usually continued to grow after the shutdown of stellar mass growth,

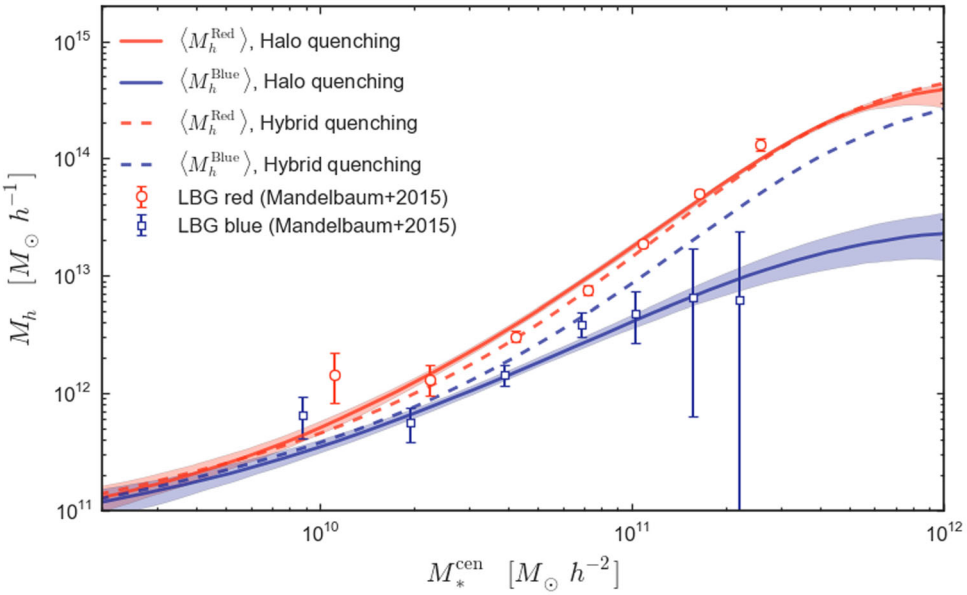

Figure 13. The average weak lensing mass of the host haloes of red and blue central galaxies as functions of stellar mass. Solid and dashed thick curves are the predictions from the halo and hybrid quenching models, respectively. The coloured bands show the 1σ uncertainties about the two mean HSMR from the fiducial halo quenching model. The circles and squares with error bars are the directly measured halo masses for the red and blue LBGs, respectively. The halo quenching model provides excellent agreement with both the red and blue LBG measurements, while the hybrid quenching model predicts a much higher halo mass for the blue LBGs.

while in star-forming systems the two often grew in sync, creating a bimodality in host halo mass between two colours without invoking halo quenching. However, this differential growth effect causes at most a factor of 2 difference in the average host halo masses, too small to explain the factor of several difference observed at the high-mass end (Quadri et al., in preparation).

The LBG experiment in Fig. 13 further demonstrates that the halo quenching model, employing halo mass as the driver for galaxy quenching, is superior to the hybrid quenching model, which relies on stellar mass to quench central galaxies. As we explained in Section 4.3, the deficiency of the hybrid model in describing the signals of the massive blue galaxies is intrinsic to the stellar mass quenching mechanism, which fails to explain the rare occurrence of blue centrals in massive clusters. The combined evidence from the BIC model comparison and the LBG experiment strongly suggests that the halo mass is the main driver for quenching the galaxies observed in SDSS.

6 PHYSICAL IMPLICATIONS OF THE CONSTRAINTS ON HALO QUENCHING MODEL

With the halo quenching model being established as the more viable scenario, we now focus back on the physical implications of our constraints on halo quenching.

6.1 Uniform characteristic halo masses for quenching centrals and satellites

Although the halo quenching formula for centrals and satellites are decoupled in the analysis, our fiducial constraint none the less recovers two very similar characteristic halo masses ($M_{\rm h}^{\rm qc}$ and $M_{\rm h}^{\rm qs}$) for both species at around $1.5\times10^{12}\,h^{-1}\,{\rm M}_\odot$. It is very tempting to associate this uniform quenching mass scale for both central and satellites to $M_{\rm shock}$, the critical halo mass responsible for the turning-on of shock heating. Analytical calculations and hydrodynamic simulations both favour an $M_{\rm shock}$ of $\sim$ few times $10^{12}\,h^{-1}\,{\rm M}_\odot$ (Birnboim & Dekel 2003; Kereš et al. 2005; Dekel & Birnboim 2006), providing one of the most plausible explanation for the similar values of $M_{\rm h}^{\rm qc}$ and $M_{\rm h}^{\rm qs}$ derived statistically in our analyses.

Conservatively speaking, even if the similarity between our inferred characteristic halo masses and $M_{\rm shock}$ were coincidental, the consistency between $M_{\rm h}^{\rm qc}$ and $M_{\rm h}^{\rm qs}$ still indicates that the quenching of centrals and satellites are somewhat coupled, most likely driven by processes that are both tied to the potential well of the haloes. For instance, the supermassive black holes (SMBHs) could provide the thermal or mechanical feedback required to stop the halo gas from cooling and feeding the satellites (Di Matteo, Springel & Hernquist 2005; Croton et al. 2006; Somerville et al. 2008), while regulating the growth of the central galaxies (Ferrarese & Merritt 2000; Gebhardt et al. 2000; Tremaine et al. 2002). Hopkins et al. (2007) suggested that the SMBH mass is largely determined by the depth of the potential well in the central regions of the system, which precedes the assembly of halo mass, i.e. the maximum circular velocity is already half the present-day value by the time the halo has accreted only two per cent of its final mass (Bosch et al. 2014).

6.2 Implications for satellite quenching and galactic conformity

The halo quenching of satellites has a slower transition across $10^{12}\,h^{-1}\,{\rm M}_\odot$ than that of the central galaxies (left-hand panel of Fig. 12). This rules out the possibility that halo quenching does not distinguish between centrals and satellites. As mentioned in the Introduction, even in the hot halo quenching scenario where gas cooling of centrals and satellites were equally inhibited, the satellites might experience significant delays in their quenching, due to a shorter exposure to the hot halo and/or a spell of star formation from pockets of cold gas they carried across the virial radius of the larger, hotter halo (van den Bosch et al. 2008; Simha et al. 2009; Wetzel,

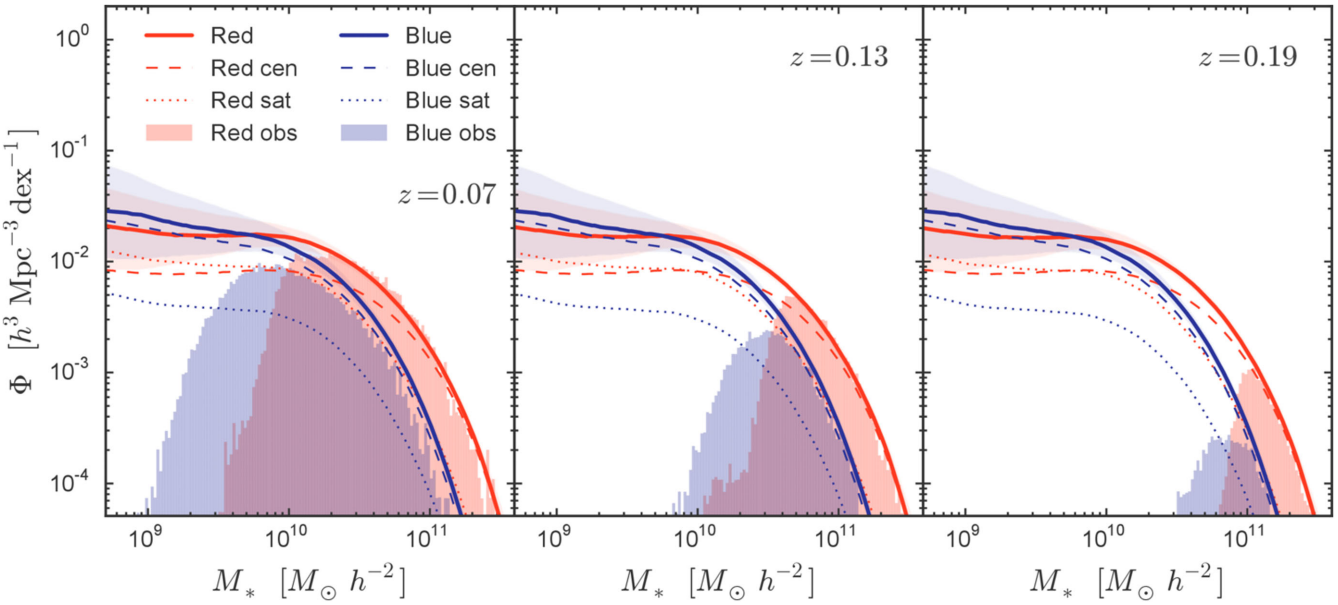

Figure 14. Comparison between the colour-segregated SMFs predicted by the best-fitting halo quenching model (thick solid curves with shaded 1σ uncertainty bands) and that observed in the SDSS (shaded histograms) at three different redshifts (from left to right: $z = 0.07$, 0.13, and 0.19). Within each colour, the total SMF is further decomposed into the central (dashed) and satellite (dotted) contributions. The excellent agreement between the predicted and observed SMFs for both the red and blue populations is not a result of fitting, but the consistency between galaxy abundance and their clustering and lensing signals predicted by our `iHOD` analysis using the halo quenching model within ΛCDM. Note that above $10^{11}\,h^{-2}\,\mathrm{M}_{\odot}$, the observed SMF in the lowest redshift bin (left; $z = 0.07$) is subject to slight incompleteness while the other two are relatively complete, hence a slight deviation from our best-fitting prediction.

Tinker & Conroy 2012; Wetzel et al. 2013). Additionally, recent observations suggest that other processes like pre-processing during infall (Haines et al. 2015), strangulation (Peng et al. 2015), and ram pressure stripping (Muzzin et al. 2014) are all at play, contributing to the satellite quenching trend with halo mass. The inefficiency of satellite quenching is also seen in dwarf galaxies below the stellar mass scale we probed here (Wheeler et al. 2014).

Another interesting aspect of the halo quenching scenario is that it may help explain galactic conformity, i.e. the observed correlation between colours of the central galaxies and their surrounding satellites (Weinmann et al. 2006; Knobel et al. 2015), because the quiescent pairs of centrals and satellites are quenched by the common haloes they share. However, this halo quenching-induced conformity only occurs among central-satellite pairs at fixed M_* of the centrals. To explain the galactic conformity observed at fixed $M_{\rm h}$, there either has to be a substantial scatter between observed $M_{\rm h}$ and true $M_{\rm h}$, or a secondary process that couples the quenching of centrals and satellites within the same halo. For instance, haloes formed earlier with higher concentration may be more likely to host quenched pairs of centrals and satellites than their younger and less concentrated counterparts at the same $M_{\rm h}$ (Paranjape et al. 2015). Galactic conformity does not appear when the centrals and satellites were quenched independently, e.g. in the hybrid quenching scenario.

The combination of this intra-halo conformity and the correlation between clustering bias and halo mass, could potentially explain the inter-halo conformity observed in the galaxy marked correlation statistics (Skibba et al. 2006; Cohn & White 2014) and hinted by galaxy pairs in the local volume ($z < 0.03$; see Kauffmann et al. 2013), although a secondary quenching induced by either formation time or halo concentration at fixed $M_{\rm h}$ may be required (Paranjape et al. 2015). We will explore the conformity prediction of the halo quenching model in the upcoming third paper of this series.

6.3 Colour-segregated SMFs of central and satellite galaxies

As another consistency check of our analysis, Fig. 14 shows the underlying SMFs of the red and blue galaxies predicted by our best-fitting halo quenching model at three redshifts ($z = 0.07$, 0.13, 0.19). In each panel, the shaded bands are the 1σ uncertainties on the predicted SMFs, and the shaded histograms show the observed SMFs of each colour, i.e. direct galaxy number counts at each redshift without the $1/V_{\rm max}$ weighting. The SMF of each colour is further decomposed into contributions from the central (dashed) and satellite (dotted) galaxies. The observed SMF at $z = 0.07$ (left-hand panel) is more incomplete at the high M_* end due to photometric confusions about bright sources in SDSS (see Paper I for details), therefore lying further below our predictions than that at the two higher redshifts. Since the observed SMFs are not used as input data to the constraints, the excellent agreement between our predictions and the direct number counts on the high stellar mass end is very encouraging – it demonstrates the great consistency between the three key observables (i.e. galaxy clustering, g–g lensing, and the SMFs) measured for the two coloured populations and that predicted by the best-fitting halo quenching model within the `iHOD` framework. Furthermore, the successful recovery of the red and blue SMFs means that the halo quenching model naturally recovers the stellar mass quenching trend observed in P10.

The best-fitting halo quenching also provides concrete prediction for the conditional SMF of the red and blue satellite galaxies, shown on the left- and right-hand panels in Fig. 15, respectively. The conditional SMF is defined as the average number of satellites per dex in stellar mass at fixed halo mass, $\langle \mathrm{d}N_{\rm sat}/\mathrm{d}\lg M_*|M_{\rm h}\rangle$, the integration of which over M_* gives the commonly used satellite HOD, i.e. the average number of galaxies per halo above some stellar mass limit $\langle N_{\rm sat}(M_*{>}M_*^{\rm lim})|M_{\rm h}\rangle$. In each panel, the solid curves are the red/blue satellite SMFs within haloes of six different

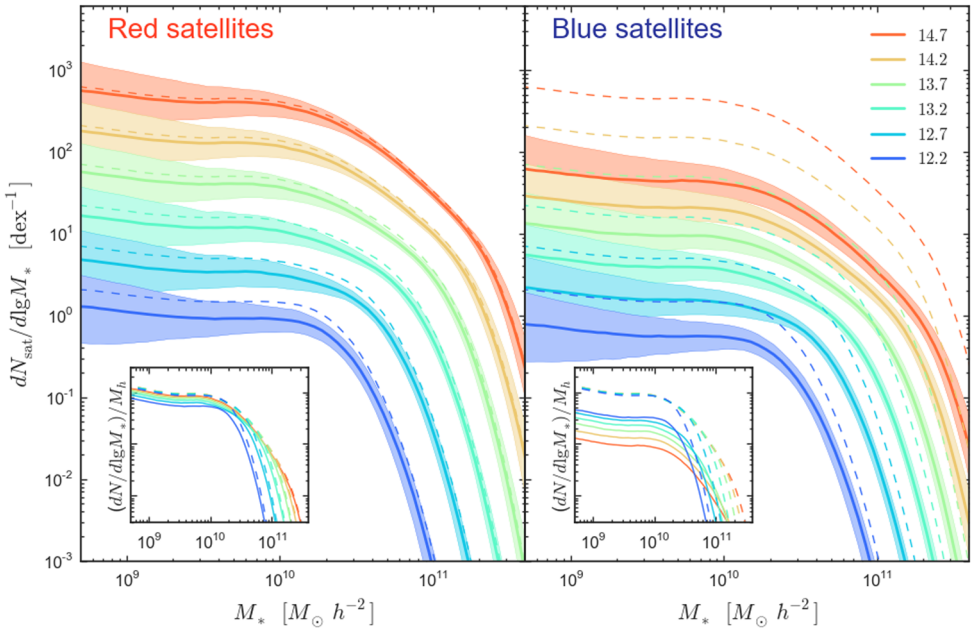

Figure 15. Red (left) and blue (right) satellite SMFs conditioned at six different host halo masses, predicted by the best-fitting halo quenching model. For each conditional SMF curve, the shaded band indicates the 1σ uncertainty and the accompanying dashed curve indicate the conditional SMF of the combined red and blue satellites. The two inset panels show similar sets of curves as the main panel, but normalized by the mass of the corresponding host haloes.

masses ($\geqslant M_{\rm h}^{\rm qs}$), with their 1σ uncertainties indicated by the shaded bands. The dashed curves are the same in both panels, showing the two-colour combined satellite SMFs. Each inset panel shows the same set of curves as in the main panel, but each normalized by the corresponding halo mass. Overall, the satellite population above $M_{\rm h}^{\rm qs}$ is dominated by the red galaxies, and the number of blue satellite galaxies per halo mass decreases with increasing halo mass due to progressively stronger halo quenching effect, while the total number of satellites per halo mass remains roughly constant.

7 COMPARISON TO PREVIOUS WORKS

Our quenching model is fundamentally different from previous studies of the link between galaxy colours and the underlying dark matter haloes. Here, we compare our best-fitting quenching model to the two main alternative methods. One is the separate red and blue galaxy modelling using traditional HOD methods (Section 7.1), and the other is based on a modified abundance matching scheme, i.e. the age-matching model (Section 7.2). We summarize the comparison between our result and the previous studies in Fig. 16, which zooms in on the stellar mass range of Fig. 13 that has the maximum model discriminating power. In addition to the LBG weak lensing masses shown in Fig. 13, we also include the average halo mass of the red and blue centrals measured from satellite kinematics by More et al. (2011, also see Conroy et al. 2007; Wojtak & Mamon 2013). We note that although the various constraints and measurements shown in Fig. 13 assumed slightly different cosmologies, the uncertainties due to cosmology are usually much smaller than the statistical errors, and the strong bimodality (or the lack thereof) in the host halo mass between red and blue is independent of any changes in cosmology. We will come back to Fig. 16 frequently and discuss individual comparisons in detail below.

7.1 Comparison to traditional HOD models

The most straightforward way to model the red and blue split of galaxy observables traditionally is to infer the HODs of the overall and the red galaxies first, and subtract the two to derive the HOD of the blue. This approach guarantees the consistency between the three sets of HODs, but lacks the flexibility in the treatment of the red fraction for describing the full ranges of behaviours seen in the data (Zehavi et al. 2005, 2011). A more comprehensive method is to treat the two colours separately, by prescribing independent HODs for the two and an 1D overall quenched fraction as a function of halo mass, as done recently in Tinker et al. (2013) and Rodríguez-Puebla et al. (2015).

As mentioned in the Introduction, there are two main differences between our approach and the methods of Tinker et al. (2013) and Rodríguez-Puebla et al. (2015). First, our quenching analysis employs only four more parameters to explain the split into red and blue galaxies, while the traditional methods require doubling of the number of parameters used for the overall population (e.g. 23 parameters in the Rodríguez-Puebla et al. 2015 analysis, and 27 in Tinker et al. 2013). Our four parameters are also more physically meaningful because they can be directly related to the average quenching action, as we discussed in Section 6. Secondly, our quenching model describes the bimodality of galaxy occupation statistics in a mathematically consistent manner – the overall galaxy HOD is recovered by summing the inferred red and blue HODs. The other two methods do not benefit from this consistency, making the connection between the colour-segregated HODs and the overall HOD hard to interpret. Naively one might think that our model is a subset of the traditional HOD models that appear more flexible by fitting red and blue separately. However, since the combination of the two separate coloured SHMRs derived in the traditional methods usually does not fit the overall galaxy population while our `iHOD` quenching models do, the two methods

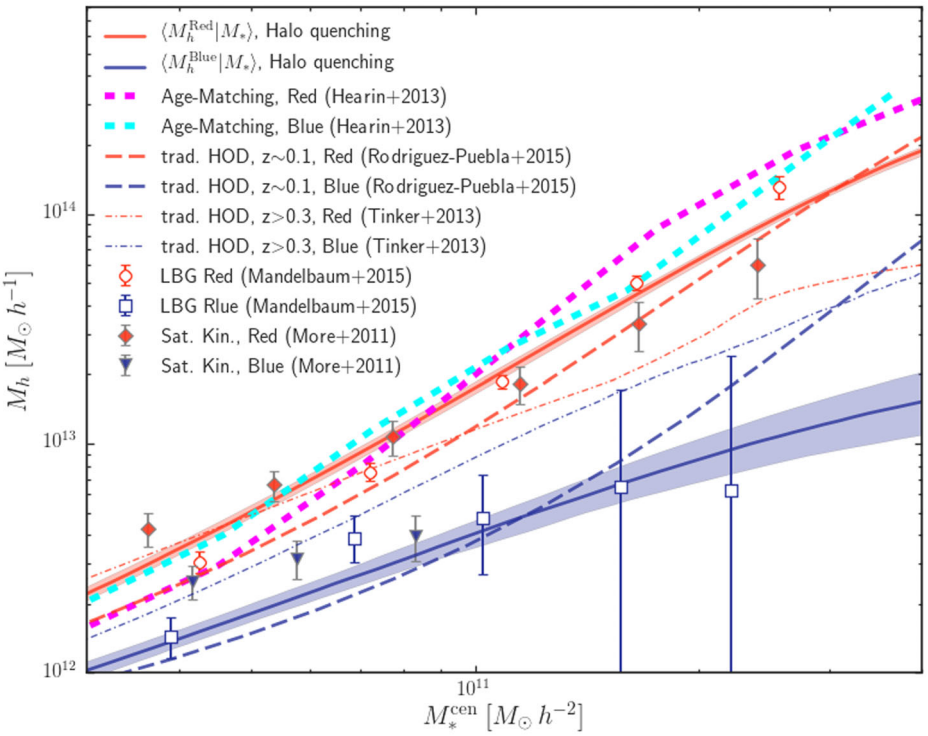

Figure 16. Comparison of various predicted and measured $\langle M_{\rm h}|M_*\rangle$ relations for the red and blue galaxies. Thick solid curves with shaded bands are the predictions from our best-fitting halo quenching model, while thick short-dashed curves are predicted from the age-matching model in Hearin et al. (2014). Long-dashed curves are the traditional HOD predictions from the Rodríguez-Puebla et al. (2015) method which derived the red and blue relations from SDSS separately, while dot–dashed curves are from Tinker et al. (2013) for COSMOS galaxies at $z > 0.3$. Filled and open symbols with error bars are the average halo masses measured from the satellite kinematics of galaxy groups (More et al. 2011) and the weak lensing of LBGs (Mandelbaum et al. 2015), respectively.

are fundamentally different descriptions of the red and blue galaxy populations.

It is worth nothing that the constraints drawn from traditional HOD modelling are unaffected by its incapability of including stellar mass quenching, which is subdominant compared to halo mass quenching. However, as we discussed in Section 6.2, the halo quenching of satellites has a much slower transition across the critical halo mass than that of centrals. Therefore, the lack of separate treatments for the red fractions in centrals and satellites remains an important issue in those traditional HOD models.

Using the SMFs, galaxy clustering, and g–g lensing within the COSMOS survey, Tinker et al. (2013) derived the SHMRs of active and quiescent galaxies over the redshift range between 0.2 and 1.0. The active/quiescent classification is based on their separation on the optical versus near-IR colour space, which is better at distinguishing dusty and SF objects than using the optical colours alone. Employing the global HOD framework of Leauthaud et al. (2012),[2] they applied independent central galaxy SHMR and satellite HODs to the two colours, and assumed a non-parametric form for the red fraction as a function of halo mass using as a spline-interpolated function through five pivotal halo masses. For the lowest redshift bin in their analysis ($z \sim 0.36$), the derived red and blue SHMRs are very similar below $M_{\rm h} \sim 10^{13}\, h^{-1}\, {\rm M}_\odot$, but strongly diverge on the high $M_{\rm h}$ end, with the red centrals having a lower average M_* than the blue ones at fixed $M_{\rm h}$. This divergence is equivalent to having a stellar mass quenching at fixed halo mass, albeit in the opposite direction of the stellar mass quenching trend observed in P10. The $\langle M_{\rm h}|M_*\rangle$ relation reveals a similar trend that our halo quenching model predicts, with the red central galaxies residing in more massive haloes than the blue centrals, but the predicted difference between the red and blue amplitudes is much smaller (dot–dashed curves).

Rodríguez-Puebla et al. (2015) derived the separate SHMRS of red and blue galaxies using the combination of galaxy clustering and SMFs in SDSS. Instead of using the g–g lensing as an input, they employed the SMFs measured for the centrals and satellites separately within each colour, using the SDSS group catalogue constructed by Yang et al. (2012). Unlike Tinker et al. (2013), they assumed a parametric form for the red fraction as a function of halo mass. The long-dashed curves in Fig. 16 are the $\langle M_{\rm h}|M_*\rangle$ relations inferred by Rodríguez-Puebla et al. (2015), showing good agreement with our predictions as well as the measurements from satellite kinematics and LBG weak lensing. The higher amplitude of $\langle M_{\rm h}|M_*\rangle$ on $M_*{>}2 \times 10^{11}\, h^{-2}\, {\rm M}_\odot$ is driven by the extrapolation of the parametric relation from lower M_* rather than data. Although Rodríguez-Puebla et al.'s and our analyses make use of the same galaxies in SDSS, the two constraints are derived using independent methods and different measurements. Hence, the good degree of consistency between these results is non-trivial.

7.2 Comparison to the age-matching model

An alternative way to describe the colour dependence of galaxy clustering and g–g lensing is to extend the SHAM model to allow

[2] The parametrization of the `iHOD` framework in Paper I is also heavily based on the Leauthaud et al. (2012) framework, but is fundamentally different in the treatment of sample completeness and signal calculation; see Paper I for a detailed comparison between the two frameworks.

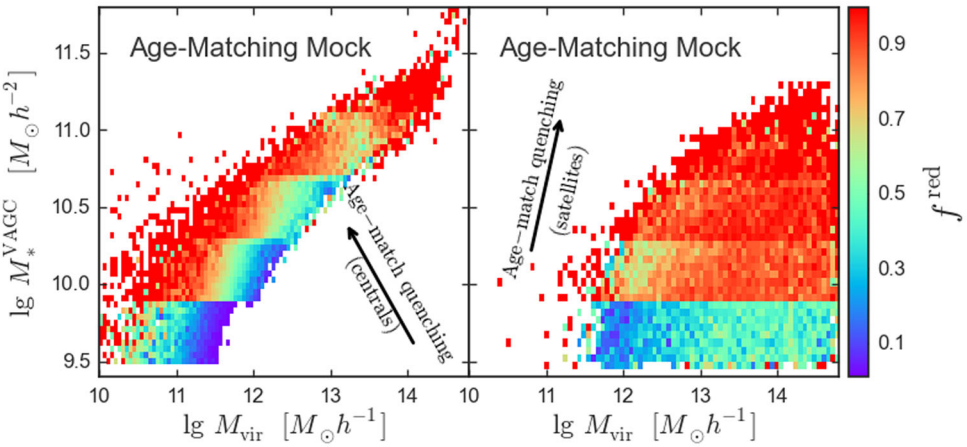

Figure 17. Similar to Fig. 4, but for the quenching within the age-matching model. The colour map shows the red galaxy fraction as a function of both stellar mass and halo mass for the central (left) and satellite (right) galaxies in the age-matching mock catalogue. Unlike other figures in this paper, we use the VAGC stellar mass and the halo virial mass used in the original age-matching catalogue. Clearly, the centrals exhibit strong stellar mass quenching trend, with a secondary formation-time (i.e. age) quenching trend at fixed M_*, while the satellites are dominated by stellar mass quenching. The reversed halo quenching due to age-matching explains the overprediction of host halo mass for blue centrals.

a secondary matching between galaxy colour and subhalo formation time, i.e. the so-called age-matching model (Hearin & Watson 2013). In particular, after the usual mapping between subhalo mass (using $v_{\rm peak}$ as a proxy; see Reddick et al. 2013) and galaxy stellar mass, the age-matching method rank-orders the characteristic redshift $z_{\rm starve}$ of those subhaloes at fixed M_* and matches the galaxies hosted by older subhaloes to redder colours, while maintaining the observed colour distribution of galaxies at that M_*. For most of the centrals (below $10^{11}\,h^{-2}\,{\rm M}_\odot$), $z_{\rm starve}$ is equivalent to the formation redshift of the subhaloes $z_{\rm form}$, but at very high M_* it is dominated by $z_{\rm char}$, the first epoch at which halo mass exceeds $10^{12}\,h^{-1}\,{\rm M}_\odot$. The age-matching method roughly reproduces the colour and stellar mass dependences of the clustering and g–g lensing signals (Hearin et al. 2014).

In the language of statistical quenching, the age-matching method describes the red fraction as

$$f_{\rm cen}^{\rm red}(M_*,\,M_{\rm h}) = f_{\rm cen}^{\rm red}(M_*) \times m(z_{\rm starve}(M_{\rm h})|M_*), \tag{36}$$

where $m(z_{\rm starve}|M_*)$ is determined by the matching between $z_{\rm starve}$ and colour at fixed M_*, and $z_{\rm starve}(M_{\rm h})$ is the formation time versus halo mass relation. Therefore, the age-matching process to first order assumes a stellar mass quenching, as the colour-matching step is done in bins of M_* regardless of halo mass (the first term on the RHS of equation 36), while the secondary quenching is via formation time (the second term on the RHS of equation 36). The combined quenching effect is best illustrated in Fig. 17, where we show the distribution of red fraction on the M_*–$M_{\rm h}$ diagram for centrals (left) and satellites (right), calculated from the age-matching mock catalogue generated by Hearin et al. (2014, with the original NYU-VAGC stellar mass and halo virial mass). For the centrals, the stellar mass quenching along the vertical axis dominates, albeit in a discrete fashion due to the binning artefact. At fixed stellar mass, since halo age is a decreasing function of halo mass (i.e. ${\rm d}z_{\rm starve}/{\rm d}M_{\rm h} < 0$; cf. fig. 12 in Wechsler et al. 2002), the secondary quenching direction of the age-matching method is a *reversed* halo quenching, with bluer centrals occupying younger, thus more massive haloes. Therefore, according to the discussion in Section 4.3, we anticipate the 'age-quenching' to predict $\langle M_{\rm h}|M_*\rangle$ relations that are similar to the hybrid quenching model, i.e. a weak segregation in the host halo mass between the red and blue centrals.

The magenta and cyan dotted curves in Fig. 16 indicate the $\langle M_{\rm h}|M_*\rangle$ relations of the red and blue galaxies, measured from the age-matching mock catalogue produced by Hearin et al. (2014, with both M_* and $M_{\rm h}$ converted to our mass units and definitions). Compared to the halo mass measured from satellite kinematics and LBG weak lensing, the age-matching mock heavily overpredicts the amplitude of the relation for the blue galaxies, making the red and blue centrals occupy haloes of similar mass, despite the systematic difference in their formation times. The tiny difference in the average halo mass between the two colours (below 0.2 dex over all mass scales) also switches sign across $M_*{\sim}10^{11}\,h^{-2}\,{\rm M}_\odot$, where the indicator for $z_{\rm starve}$ switches from $z_{\rm form}$ to $z_{\rm char}$. In particular, below $10^{11}\,h^{-2}\,{\rm M}_\odot$, age-matching predicts that the blue centrals live in more massive haloes than the red centrals, due to the anticorrelation between $z_{\rm form}$ and halo mass – at any given M_* the younger haloes that are assigned bluer centrals are also more massive. However, this increase of halo mass with bluer colour of the centrals is in the opposite direction compared to the observations. Above $10^{11}\,h^{-2}\,{\rm M}_\odot$, the characteristic redshift $z_{\rm starve}$ is dominated by $z_{\rm char}$, which is more positively correlated with halo mass and assigns redder centrals to more massive systems. The overall disagreement between the $\langle M_{\rm h}|M_*\rangle$ predicted by the age-matching method and that measured from satellite kinematics and LBG weak lensing indicates that the stellar mass quenching assumed in age-matching is not adequate, and the secondary formation-time quenching *at fixed stellar mass* is strongly disfavoured by the observations.

Another difference between the age-matching model and the quenching models considered in this paper is that, by choosing formation time as a quenching indicator, the age-matching model exhibits the maximum level of galactic assembly bias (Zentner, Hearin & van den Bosch 2014), which is absent in our quenching models, by construction. However, while the 'halo assembly bias', namely, the dependence of halo properties on the formation history, is clearly detected in cosmological simulations (Sheth & Tormen 2004; Gao, Springel & White 2005; Harker et al. 2006; Wechsler et al. 2006; Zhu et al. 2006; Hahn et al. 2007; Jing, Suto & Mo 2007; Croft et al. 2012), whether it left a significant imprint on the observed galaxy properties is still in debate (Berlind et al. 2006; Blanton & Berlind 2007; Wang et al. 2013; Lin et al. 2015; Miyatake et al. 2016). In mock galaxy catalogues constructed from

semi-analytical models and hydrodynamic simulations, galaxy clustering exhibits an assembly bias of at most ∼10 per cent (Yoo et al. 2006; Croton, Gao & White 2007; Zu et al. 2008; Mehta 2014).[3] The great success of the halo quenching model in quantitatively explaining the clustering and weak lensing of the red and blue galaxies, while simultaneously recovering their respective SMFs and average halo masses, strongly suggests that the halo quenching is the dominant process in shaping the distribution of galaxy colours observed in SDSS, and that any impact of the galactic assembly bias should be a secondary effect, in the form of, e.g. a formation-time (or concentration) quenching *at fixed halo mass* proposed by Paranjape et al. (2015).

Finally, the physical interpretation of the halo quenching model (as discussed in Section 6.1) relates the quenching of galaxies to the capability of the host haloes to either heat the incoming gas or keep the hot gas from cooling, and the sharp transition of this capability across some critical halo mass is the key to explain the strong bimodality in galaxy colours. This physical picture is fundamentally different from the physical motivation of age-matching, in which the galaxy quenching is strictly tied to the dark matter accretion history of haloes at fixed M_*. However, the average halo accretion history is a smooth function of cosmic time, therefore showing no bimodality in the formation time of haloes (Zhao et al. 2009). In addition, the connection between dark matter accretion and SFR is very complex. Using a suite of high-resolution hydrodynamic simulations, Faucher-Giguère, Kereš & Ma (2011) showed that the cold gas accretion rate, which is more directly related to star formation, is in general not a simple universal factor of the dark matter accretion rate, and that baryonic feedback can cause SFRs to deviate significantly from the external gas accretion rates.

8 CONCLUSIONS

We develop a novel method to identify the dominant driver of galaxy quenching in the local Universe, using the galaxy clustering and g–g lensing of red and blue galaxies observed in SDSS. The method extends the powerful `iHOD` framework developed in Paper I by introducing two quenching models: (1) a *halo* quenching model in which the average probability of a galaxy being quenched depends solely on the main halo mass, but in separate manners for centrals and satellites; and (2) a *hybrid* quenching model in which the quenching probability of all galaxies depends on their stellar mass, with the satellite having an extra dependence on the host halo mass.

The two quenching models predict distinctive 2D distributions of red galaxy fractions on the M_*–$M_{\rm h}$ plane, resulting in different patterns through which red and blue galaxies populate dark matter haloes. We then predict the clustering and g–g lensing signals of the red and blue galaxies from these two galaxy occupation patterns using the `iHOD` framework and compare them to the measurements from SDSS. Most importantly, the flexibility of `iHOD` allows us to include ∼80 per cent more galaxies in the analysis than the traditional HOD method, greatly enhancing our capability of statistically distinguishing the two quenching models.

We find that the halo quenching model provides better descriptions of the bimodality in the clustering and lensing of observed galaxies than the hybrid quenching models, mainly due to the significantly improved fit to massive blue galaxies. We further identify that the average host halo mass of the massive blue centrals provides the most discriminating power in testing viable quenching models – models with halo mass quenching generally predict a much stronger segregation in the average host halo mass ($\langle M_{\rm h}|M_*\rangle$) between the red and blue than the ones without (Fig. 13).

Therefore, by comparing the $\langle M_{\rm h}|M_*\rangle$ predicted by various quenching models, including the age-matching model, to that measured directly from the satellite kinematics of galaxy clusters and the weak lensing of LBGs, we confirm that the best-fitting halo quenching model provides excellent agreement with the two observational measurements, while models that rely on stellar mass (e.g. the hybrid quenching model and the age-matching method) fail to predict the halo mass of the blue central galaxies (Fig. 16). Furthermore, the formation time-quenching at fixed M_* prescribed in the age-matching method creates a *reversed* halo quenching trend, therefore placing blue and red centrals of the same M_* into higher and lower mass haloes, respectively. This trend is strongly disfavoured by the observations where at any given M_* redder centrals on average occupy more massive haloes. Our findings indicate that any viable abundance matching scheme for assigning galaxy colours has to reproduce the observed strong bimodality in host halo mass between red and blue centrals, especially at the high-mass end.

The derived characteristic halo masses of the central and satellite quenching have very similar values around $1.5 \times 10^{12}\,h^{-1}\,{\rm M}_\odot$, suggesting that a uniform halo quenching process is operating on both the centrals and satellites. The derived characteristic halo mass can be interpreted by the canonical halo quenching theory, which predicts a critical halo mass of $M_{\rm shock} \sim 10^{12}\,h^{-1}\,{\rm M}_\odot$. Above this critical mass, the virial shock is able to prevent star formation by heating the infalling gas to high temperatures, whereas below $M_{\rm shock}$ the halo quenching rapidly turns off, creating strong bimodality in both the colour and the spatial distribution of galaxies. The pace at which the halo mass quenching operates, however, appears to be faster for the centrals than for the satellite galaxies, which are quenched in a more delayed and prolonged fashion. It would be interesting to combine our analysis of the local Universe with observations at higher redshifts (Brodwin et al. 2013; Kawinwanichakij et al. 2014; Lin et al. 2016), where cold streams were strong enough to penetrate through the hot media and form massive blue discs in haloes above $M_{\rm shock}$ (Dekel & Birnboim 2006).

In the future, we anticipate that the `iHOD` halo quenching model, which accurately explains the spatial clustering, g–g lensing, and SMFs of the red and blue galaxies, will provide an important baseline model for explaining an even wider range of observed galaxy properties with ever-growing precision. In the near term, i.e. the upcoming Paper III in this series, we plan to generate realistic colour-segregated galaxy mock catalogues using the constraints inferred from the halo quenching model analysis in this paper, and make a comprehensive comparison with the observed galaxies to look for potential signatures of any secondary quenching processes, e.g. due to formation time (Paranjape et al. 2015) and galaxy compactness (Woo et al. 2015) at fixed halo mass.

ACKNOWLEDGEMENTS

We thank Aldo Rodríguez-Puebla, Surhud More, and Jeremy Tinker for kindly providing their measurements. We also thank Hung-Jin Huang for useful discussions. We thank David Weinberg and Zheng Zheng for carefully reading an earlier version of the manuscript and

[3] Some hydrodynamic simulations predict much higher assembly bias effect on clustering (∼20 per cent), which cannot be captured by any current abundance matching models (Chaves-Montero et al. 2015).

for giving detailed comments and suggestions that have greatly improved the manuscript. YZ and RM acknowledge the support by the Department of Energy Early Career Program, and the Alfred P. Sloan Fellowship program.

REFERENCES

Abazajian K. N. et al., 2009, ApJS, 182, 543
Baldry I. K., Balogh M. L., Bower R. G., Glazebrook K., Nichol R. C., Bamford S. P., Budavari T., 2006, MNRAS, 373, 469
Balogh M. L., Morris S. L., 2000, MNRAS, 318, 703
Benson A. J., 2010, Phys. Rep., 495, 33
Berlind A. A., Kazin E., Blanton M. R., Pueblas S., Scoccimarro R., Hogg D. W., 2006, preprint (astro-ph/0610524)
Bernardi M., Shankar F., Hyde J. B., Mei S., Marulli F., Sheth R. K., 2010, MNRAS, 404, 2087
Binney J., 1977, ApJ, 215, 483
Binney J., 2004, MNRAS, 347, 1093
Birnboim Y., Dekel A., 2003, MNRAS, 345, 349
Blanton M. R., Berlind A. A., 2007, ApJ, 664, 791
Blanton M. R., Eisenstein D., Hogg D. W., Schlegel D. J., Brinkmann J., 2005, ApJ, 629, 143
Bosch F. C. v. d., Jiang F., Hearin A., Campbell D., Watson D., Padmanabhan N., 2014, MNRAS, 445, 1713
Brodwin M. et al., 2013, ApJ, 779, 138
Bruzual G., Charlot S., 2003, MNRAS, 344, 1000
Cattaneo A., Dekel A., Devriendt J., Guiderdoni B., Blaizot J., 2006, MNRAS, 370, 1651
Chabrier G., 2003, PASP, 115, 763
Chaves-Montero J., Angulo R. E., Schaye J., Schaller M., Crain R. A., Furlong M., 2015, preprint (arXiv:1507.01948)
Cohn J. D., White M., 2014, MNRAS, 440, 1712
Conroy C., Wechsler R. H., Kravtsov A. V., 2006, ApJ, 647, 201
Conroy C. et al., 2007, ApJ, 654, 153
Croft R. A. C., Matteo T. D., Khandai N., Springel V., Jana A., Gardner J. P., 2012, MNRAS, 425, 2766
Croton D. J. et al., 2006, MNRAS, 365, 11
Croton D. J., Gao L., White S. D. M., 2007, MNRAS, 374, 1303
Daddi E. et al., 2007, ApJ, 670, 156
Dekel A., Birnboim Y., 2006, MNRAS, 368, 2
Di Matteo T., Springel V., Hernquist L., 2005, Nature, 433, 604
Faucher-Giguère C.-A., Kereš D., Ma C.-P., 2011, MNRAS, 417, 2982
Feldmann R. et al., 2006, MNRAS, 372, 565
Ferrarese L., Merritt D., 2000, ApJ, 539, L9
Gabor J. M., Davé R., 2015, MNRAS, 447, 374
Gao L., Springel V., White S. D. M., 2005, MNRAS, 363, L66
Gebhardt K. et al., 2000, ApJ, 539, L13
Gunn J. E., Gott J. R., III, 1972, ApJ, 176, 1
Hahn O., Porciani C., Carollo C. M., Dekel A., 2007, MNRAS, 375, 489
Haines C. P. et al., 2015, ApJ, 806, 101
Harker G., Cole S., Helly J., Frenk C., Jenkins A., 2006, MNRAS, 367, 1039
Hearin A. P., Watson D. F., 2013, MNRAS, 435, 1313
Hearin A. P., Watson D. F., Becker M. R., Reyes R., Berlind A. A., Zentner A. R., 2014, MNRAS, 444, 729
Hirata C., Seljak U., 2003, MNRAS, 343, 459
Hopkins P. F., Hernquist L., Cox T. J., Robertson B., Krause E., 2007, ApJ, 669, 45
Huang S., Gu Q.-S., 2009, MNRAS, 398, 1651
Jing Y. P., Suto Y., Mo H. J., 2007, ApJ, 657, 664
Kass R. E., Raftery A. E., 1995, J. Am. Stat. Assoc., 90, 773
Katz N., Keres D., Dave R., Weinberg D. H., 2003, in Rosenberg J. L., Putman M. E., eds, Astrophysics and Space Science Library Vol. 281, The IGM/Galaxy Connection. The Distribution of Baryons at z=0. Kluwer, Dordrecht, p. 185
Kauffmann G. et al., 2003, MNRAS, 341, 33
Kauffmann G., Li C., Zhang W., Weinmann S., 2013, MNRAS, 430, 1447
Kawinwanichakij L. et al., 2014, ApJ, 792, 103
Kereš D., Katz N., Weinberg D. H., Davé R., 2005, MNRAS, 363, 2
Kereš D., Katz N., Fardal M., Davé R., Weinberg D. H., 2009, MNRAS, 395, 160
Knobel C., Lilly S. J., Woo J., Kovač K., 2015, ApJ, 800, 24
Landy S. D., Szalay A. S., 1993, ApJ, 412, 64
Leauthaud A. et al., 2012, ApJ, 744, 159
Lin Y.-T., Mandelbaum R., Huang Y.-H., Huang H.-J., Dalal N., Diemer B., Jian H.-Y., Kravtsov A., 2015, preprint (arXiv:1504.07632)
Lin L. et al., 2016, ApJ, 817, 97
Mandelbaum R. et al., 2005, MNRAS, 361, 1287
Mandelbaum R., Hirata C. M., Leauthaud A., Massey R. J., Rhodes J., 2012, MNRAS, 420, 1518
Mandelbaum R., Slosar A., Baldauf T., Seljak U., Hirata C. M., Nakajima R., Reyes R., Smith R. E., 2013, MNRAS, 432, 1544
Mandelbaum R., Wang W., Zu Y., White S., Henriques B., More S., 2015, MNRAS, in press, preprint (arXiv:1509.06762)
Mehta K. T., 2014, PhD thesis, Univ. Arizona
Miyatake H., More S., Takada M., Spergel D. N., Mandelbaum R., Rykoff E. S., Rozo E., 2016, Phys. Rev. Lett., 116, 1301
More S., van den Bosch F. C., Cacciato M., Skibba R., Mo H. J., Yang X., 2011, MNRAS, 410, 210
Muzzin A. et al., 2014, ApJ, 796, 65
Nakajima R., Mandelbaum R., Seljak U., Cohn J. D., Reyes R., Cool R., 2012, MNRAS, 420, 3240
Noeske K. G. et al., 2007, ApJ, 660, L43
Paranjape A., Kovac K., Hartley W. G., Pahwa I., 2015, MNRAS, 454, 3030
Peng Y.-j. et al., 2010, ApJ, 721, 193 (P10)
Peng Y.-j., Lilly S. J., Renzini A., Carollo M., 2012, ApJ, 757, 4
Peng Y., Maiolino R., Cochrane R., 2015, Nature, 521, 192
Reddick R. M., Wechsler R. H., Tinker J. L., Behroozi P. S., 2013, ApJ, 771, 30
Reyes R., Mandelbaum R., Gunn J. E., Nakajima R., Seljak U., Hirata C. M., 2012, MNRAS, 425, 2610
Rodríguez-Puebla A., Avila-Reese V., Yang X., Foucaud S., Drory N., Jing Y. P., 2015, ApJ, 799, 130
Salim S. et al., 2007, ApJS, 173, 267
Sheth R. K., Tormen G., 2004, MNRAS, 350, 1385
Simha V., Weinberg D. H., Davé R., Gnedin O. Y., Katz N., Kereš D., 2009, MNRAS, 399, 650
Skibba R., Sheth R. K., Connolly A. J., Scranton R., 2006, MNRAS, 369, 68
Somerville R. S., Hopkins P. F., Cox T. J., Robertson B. E., Hernquist L., 2008, MNRAS, 391, 481
Speagle J. S., Steinhardt C. L., Capak P. L., Silverman J. D., 2014, ApJS, 214, 15
Strateva I. et al., 2001, AJ, 122, 1861
Tinker J. L., Leauthaud A., Bundy K., George M. R., Behroozi P., Massey R., Rhodes J., Wechsler R. H., 2013, ApJ, 778, 93
Tremaine S. et al., 2002, ApJ, 574, 740
van den Bosch F. C., Aquino D., Yang X., Mo H. J., Pasquali A., McIntosh D. H., Weinmann S. M., Kang X., 2008, MNRAS, 387, 79
Wang L., Weinmann S. M., De Lucia G., Yang X., 2013, MNRAS, 433, 515
Wang W., White S., Mandelbaum R., Henriques B., Anderson M. E., Han J., 2016, MNRAS, 456, 2301
Wechsler R. H., Bullock J. S., Primack J. R., Kravtsov A. V., Dekel A., 2002, ApJ, 568, 52
Wechsler R. H., Zentner A. R., Bullock J. S., Kravtsov A. V., Allgood B., 2006, ApJ, 652, 71
Weinmann S. M., van den Bosch F. C., Yang X., Mo H. J., 2006, MNRAS, 366, 2
Wetzel A. R., Tinker J. L., Conroy C., 2012, MNRAS, 424, 232
Wetzel A. R., Tinker J. L., Conroy C., van den Bosch F. C., 2013, MNRAS, 432, 336
Wheeler C., Phillips J. I., Cooper M. C., Boylan-Kolchin M., Bullock J. S., 2014, MNRAS, 442, 1396

Wojtak R., Mamon G. A., 2013, MNRAS, 428, 2407
Woo J. et al., 2013, MNRAS, 428, 3306
Woo J., Dekel A., Faber S. M., Koo D. C., 2015, MNRAS, 448, 237
Yang X., Mo H. J., van den Bosch F. C., Zhang Y., Han J., 2012, ApJ, 752, 41
Yoo J., Tinker J. L., Weinberg D. H., Zheng Z., Katz N., Davé R., 2006, ApJ, 652, 26
York D. G. et al., 2000, AJ, 120, 1579
Zehavi I. et al., 2005, ApJ, 630, 1
Zehavi I. et al., 2011, ApJ, 736, 59
Zentner A. R., Hearin A. P., van den Bosch F. C., 2014, MNRAS, 443, 3044
Zhao D. H., Jing Y. P., Mo H. J., Börner G., 2009, ApJ, 707, 354
Zhu G., Zheng Z., Lin W. P., Jing Y. P., Kang X., Gao L., 2006, ApJ, 639, L5
Zu Y., Mandelbaum R., 2015, MNRAS, 454, 1161 (Paper I)
Zu Y., Zheng Z., Zhu G., Jing Y. P., 2008, ApJ, 686, 41

This paper has been typeset from a TEX/LATEX file prepared by the author.