A COSMOLOGICAL FRAMEWORK FOR THE CO-EVOLUTION OF QUASARS, SUPERMASSIVE BLACK HOLES, AND ELLIPTICAL GALAXIES. I. GALAXY MERGERS AND QUASAR ACTIVITY

PHILIP F. HOPKINS,[1] LARS HERNQUIST,[1] THOMAS J. COX,[1] AND DUŠAN KEREŠ[1]

ABSTRACT

We develop a model for the cosmological role of mergers in the evolution of starbursts, quasars, and spheroidal galaxies. By combining theoretically well-constrained halo and subhalo mass functions as a function of redshift and environment with empirical halo occupation models, we can estimate where galaxies of given properties live at a particular epoch. This allows us to calculate, in an a priori cosmological manner, where major galaxy-galaxy mergers occur and what kinds of galaxies merge, at all redshifts. We compare this with the observed mass functions, clustering, fractions as a function of halo and galaxy mass, and small-scale environments of mergers, and we show that this approach yields robust estimates in good agreement with observations and can be extended to predict detailed properties of mergers. Making the simple *Ansatz* that major, gas-rich mergers cause quasar activity (but not strictly assuming they are the only triggering mechanism), we demonstrate that this model naturally reproduces the observed rise and fall of the quasar luminosity density at $z = 0-6$, as well as quasar luminosity functions, fractions, host galaxy colors, and clustering as a function of redshift and luminosity. The recent observed excess of quasar clustering on small scales at $z \sim 0.2$–2.5 is a natural prediction of our model, as mergers will preferentially occur in regions with excess small-scale galaxy overdensities. In fact, we demonstrate that quasar environments at all observed redshifts correspond closely to the empirically determined small group scale, where major mergers of $\sim L_*$ gas-rich galaxies will be most efficient. We contrast this with a secular model in which quasar activity is driven by bars or other disk instabilities, and we show that, while these modes of fueling probably dominate the high Eddington ratio population at Seyfert luminosities (significant at $z = 0$), the constraints from quasar clustering, observed pseudobulge populations, and disk mass functions suggest that they are a small contributor to the $z \gtrsim 1$ quasar luminosity density, which is dominated by massive BHs in predominantly classical spheroids formed in mergers. Similarly, low-luminosity Seyferts do not show a clustering excess on small scales, in agreement with the natural prediction of secular models, but bright quasars at all redshifts do so. We also compare recent observations of the colors of quasar host galaxies and show that these correspond to the colors of recent merger remnants, in the transition region between the blue cloud and the red sequence, and are distinct from the colors of systems with observed bars or strong disk instabilities. Even the most extreme secular models, in which all bulge (and therefore BH) formation proceeds via disk instability, are forced to assume that this instability acts before the (dynamically inevitable) mergers, and therefore predict a history for the quasar luminosity density that is shifted to earlier times, in disagreement with observations. Our model provides a powerful means to predict the abundance and nature of mergers and to contrast cosmologically motivated predictions of merger products such as starbursts and active galactic nuclei.

Subject headings: cosmology: theory — galaxies: active — galaxies: evolution — quasars: general

Online material: color figures

1. INTRODUCTION

1.1. *Motivation*

Over the past decade, observations have established that supermassive black holes likely reside in the centers of all galaxies with spheroids (e.g., Kormendy & Richstone 1995; Richstone et al. 1998; Kormendy & Gebhardt 2001) and that the properties of these black holes and their hosts are correlated. These correlations take various forms, relating the black hole mass to, e.g., the mass (Magorrian et al. 1998; McLure & Dunlop 2002; Marconi & Hunt 2003; Häring & Rix 2004), velocity dispersion (Ferrarese & Merritt 2000; Gebhardt et al. 2000; Tremaine et al. 2002), and concentration or Sersic index (Graham et al. 2001; Graham & Driver 2007) of the spheroid. Most recently, Hopkins et al. (2007c) have demonstrated that these relationships are not independent and can be understood as various projections of a black hole fundamental plane analogous to the fundamental plane for

elliptical galaxies (Dressler et al. 1987; Djorgovski & Davis 1987). The striking similarity between these two fundamental planes indicates that galaxy spheroids and supermassive black holes are not formed independently but originate via a common physical process.

Furthermore, although there may be some relatively weak evolution in the correlation between BH mass and host mass or velocity dispersion owing to changes in spheroid structural properties and internal correlations with redshift (e.g., Peng et al. 2006; Shields et al. 2003, 2006; Walter et al. 2004; Salviander et al. 2006; Woo et al. 2006; Hopkins et al. 2006e), the fundamental plane appears to be preserved (Hopkins et al. 2007c), and in any case *some* correlation exists at all redshifts. There are not, at any redshifts, bulgeless systems with large black holes or bulges without correspondingly large black holes. This empirically demonstrates that whatever process builds up black hole mass *must* trace the formation of spheroids (albeit with potentially redshift-dependent efficiency).

These connections extend to other phenomena associated with galaxies that have sometimes been interpreted as being

[1] Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138.

independent. For example, by estimating the total energy radiated by quasars, Soltan (1982) showed that nearly all the mass in supermassive black holes must have been accumulated during periods of bright quasar activity. This analysis has since been revisited on a number of occasions (Salucci et al. 1999; Yu & Tremaine 2002; Marconi et al. 2004; Shankar et al. 2004; Yu & Lu 2004), with various assumptions for quasar obscuration and bolometric corrections. Hopkins et al. (2007e) have reformulated the Soltan argument from the evolution of the *bolometric* quasar luminosity function (LF). In their analysis, Hopkins et al. combined observations of the quasar LF in a variety of wave bands with purely empirical determinations of the luminosity dependence of quasar obscuration and spectral emission to infer the bolometric quasar LF. By integrating this over luminosity and redshift, it is then possible to obtain a *model-independent* estimate of the total energy density of radiation from quasars. The cosmic black hole mass density then follows if black holes in quasars accrete with constant radiative efficiency ϵ_r (Shakura & Sunyaev 1973), by integrating $L_{\rm bol} = \epsilon_r \dot{M}_{\rm BH} c^2$. This yields a $z = 0$ black hole mass density of

$$\rho_{\rm BH}(z=0) = 4.81^{+1.24}_{-0.99} \left(\frac{0.1}{\epsilon_r} \right) h_{70}^2 \times 10^5 \; M_\odot \; {\rm Mpc}^{-3}, \quad (1)$$

consistent with estimates of $\rho_{\rm BH}(z=0)$ obtained from local bulge mass, luminosity, and velocity dispersion functions (e.g., Marconi et al. 2004; Shankar et al. 2004).

Taken together, the black hole fundamental plane and the Soltan argument imply that the common physical process that produces galaxy spheroids and supermassive black holes also must be responsible for triggering *most* bright quasars. Moreover, there is compelling evidence that quasar activity is preceded by a period of intense star formation in galaxy centers so that, for example, ultraluminous infrared galaxies (ULIRGs) and distant submillimeter galaxies (SMGs) would eventually evolve into quasars (Sanders et al. 1988a, 1988c; Sanders & Mirabel 1996; Dasyra et al. 2007). Essentially all sufficiently deep studies of the spectral energy distributions (SEDs) of quasar host galaxies reveal the presence of young stellar populations indicative of a recent starburst (Brotherton et al. 1999; Canalizo & Stockton 2001; Kauffmann et al. 2003; Yip et al. 2004; Jahnke et al. 2004a, 2004b; Sánchez et al. 2004; Vanden Berk et al. 2006; Barthel 2006; Zakamska et al. 2006). There further appears to be a correlation in the sense that the most luminous quasars have the youngest host stellar populations (Jahnke et al. 2004a; Vanden Berk et al. 2006) and the greatest prominence of postmerger tidal features and disturbances (Canalizo & Stockton 2001; Kauffmann et al. 2003; Hutchings et al. 2003, 2006; Hutchings 2003; Zakamska et al. 2006; Letawe et al. 2007). These observations indicate that intense starbursts must result from the same process as most quasars and supermassive black holes.

In the simplest interpretation, we seek an explanation for the various phenomena summarized above such that they result from the *same event*. There are general, theoretical requirements that any such event must satisfy. In particular, it must be fast and violent, blend together gas and stellar dynamics appropriately, and involve a supply of mass comparable to that in large galaxies. Why should this be the case?

The accepted picture for the growth of supermassive black holes is that the mass is primarily assembled by gas accretion (Lynden-Bell 1969). From the Soltan argument, we know that this mass must be gathered in a time comparable to the lifetimes of bright quasars, which is similar to the Salpeter (1964) time $\sim 10^{7.5}$ yr, for black holes accreting at the Eddington rate. Inde-

pendent limits (Martini 2004 and references therein) from quasar clustering, variability, luminosity function evolution, and other methods demand a *total* quasar lifetime (i.e., duration of major growth for a given BH) of $\lesssim 10^{8.5}$ yr. In order to explain the existence of black holes with masses $\sim 10^9 \, M_\odot$, the amount of gas required is likely comparable to that contained in entire large galaxies. Thus, the process we seek must be able to deliver a galaxy's worth of gas to the inner regions of a galaxy on a relatively short timescale, $\ll 10^9$ yr.

If this event is to simultaneously build galaxy spheroids, it must involve stellar dynamics acting on a supply of stars similar to that in large galaxies because the stellar mass is ~ 1000 times larger than that of the black hole and it is believed that spheroids are assembled mainly (albeit not entirely) through dissipationless physics (i.e., the movement of stars from a circular disk to random spheroid orbits). A plausible candidate process is violent relaxation (e.g., Lynden-Bell 1967), which has been demonstrated to yield phase space distributions akin to those of elliptical galaxies through large, rapid fluctuations in the gravitational potential. Violent relaxation operates on a timescale similar to the free-fall time for self-gravitating systems, again $\ll 10^9$ yr for the bulk of the mass.

Motivated by these considerations, Hopkins et al. (2006a) developed a model in which starbursts, quasars, supermassive black hole growth, and the formation of red, elliptical galaxies are connected through an evolutionary sequence, caused by *mergers* between *gas-rich* galaxies. There is, in fact, considerable observational evidence indicating that mergers are responsible for triggering ULIRGs, SMGs, and quasars (see references in Hopkins et al. 2006a; for reviews see Barnes & Hernquist 1992; Schweizer 1998; Jogee 2006). Furthermore, the long-standing "merger hypothesis," which proposes that most elliptical galaxies formed in mergers (Toomre & Toomre 1972; Toomre 1977), is supported by the structure of known ongoing mergers (e.g., Schweizer 1992; Rothberg & Joseph 2006a, 2006b) and the ubiquitous presence of fine structures such as shells, ripples, tidal plumes, nuclear light excesses, and kinematic subsystems in ellipticals (e.g., Schweizer & Seitzer 1992; Schweizer 1996), which are signatures of mergers (e.g., Quinn 1984; Hernquist & Quinn 1987; Hernquist & Spergel 1992; Hernquist & Barnes 1991; Mihos & Hernquist 1994a).

Numerical simulations performed during the past 20 years verify that *major* mergers of *gas-rich* disk galaxies can plausibly account for these phenomena and have elucidated the underlying physics. Tidal torques excited during a merger lead to rapid inflows of gas into the centers of galaxies (Hernquist 1989; Barnes & Hernquist 1991, 1996). The amount of gas involved can be a large fraction of that in the progenitor galaxies and is accumulated on roughly a dynamical time in the inner regions, $\ll 10^9$ yr (Hernquist 1989). The resulting high gas densities trigger starbursts (Mihos & Hernquist 1994b, 1996) and feed rapid black hole growth (Di Matteo et al. 2005). Gas consumption by the starburst and dispersal of residual gas by supernova-driven winds and feedback from black hole growth (Springel et al. 2005a) terminate star formation so that the remnant quickly evolves from a blue to a red galaxy. The stellar component of the progenitors provides the bulk of the material for producing the remnant spheroid (Barnes 1988, 1992; Hernquist 1992, 1993) through violent relaxation.

The simulations also place significant constraints on the types of mergers that can initiate this sequence of events. First, a major merger is generally required in order for the tidal forces to excite a sufficiently strong response to set up nuclear inflows of gas. Although simulations involving minor mergers with mass ratios $\sim 10:1$ show that gas inflows can be excited under some circumstances (e.g., Hernquist 1989; Hernquist & Mihos 1995;

Bournaud et al. 2005), a systematic study indicates that such an outcome is limited to specific orbital geometries (J. Younger et al. 2008, in preparation) and that the overall efficiency of triggering inflows declines rapidly with increasing mass ratio. Thus, while the precise definition of a major merger in this context is blurred by the degeneracy between the mass ratio of the progenitors and the orbit of the interaction, it appears that a mass ratio ~3:1 or smaller is needed. This is further supported by observational studies (Dasyra et al. 2006; Woods et al. 2006), which find that strong gas inflows and nuclear starbursts are typically seen only below these mass ratios, despite the much greater frequency of higher mass-ratio mergers.

Second, the merging galaxies must contain a supply of *cold* gas, which in this context refers to gas that is rotationally supported, in order that the resonant response leading to nuclear inflows of gas in a merger be excited. Elliptical galaxies contain large quantities of hot, thermally supported gas, but even major mergers between two such objects will not drive the nuclear inflows of gas that fuel rapid black hole growth.

It also must be emphasized that essentially all numerical studies of spheroid kinematics find that *only* mergers can reproduce the observed kinematic properties of elliptical galaxies and "classical" bulges (Hernquist 1989, 1992, 1993; Barnes 1988, 1992; Schweizer 1992; Naab et al. 1999, 2006a, 2006b; Naab & Burkert 2003; Naab & Trujillo 2006; Bournaud et al. 2005; Jesseit et al. 2007; Cox et al. 2006b). Disk instabilities and secular evolution (e.g., bar instabilities, harassment, and other isolated modes) can indeed produce bulges, but these are invariably "pseudobulges" (Schwarz 1981; Athanassoula et al. 1983; Pfenniger 1984; Combes et al. 1990; Raha et al. 1991; Kuijken & Merrifield 1995; O'Neill & Dubinski 2003; Athanassoula 2005) with clearly distinct shapes (e.g., flattened or "peanut"-shaped isophotes), rotation properties (large v/σ), internal correlations (obeying different Kormendy and Faber-Jackson relations), light profiles (nearly exponential Sersic profiles), and colors and/or substructure from classical bulges (for a review see Kormendy & Kennicutt 2004). Observations indicate that pseudobulges constitute only a small fraction of the total mass density in spheroids (≲10%; see Allen et al. 2006; Ball et al. 2006; Driver et al. 2007), becoming a large fraction of the bulge population only for small bulges in late-type hosts (e.g., Sb/c, corresponding to typical $M_{BH} \lesssim 10^7\ M_\odot$; see Carollo et al. 1998, Kormendy & Kennicutt 2004, and references therein). Therefore, it is clear that although such processes may be important for the buildup of the smallest black hole and spheroid populations, secular evolution *cannot* be the agent responsible for the formation of most elliptical galaxies, or for the buildup of most black hole mass, or the triggering of bright quasar activity.

We are thus led to suggest a generalization of the merger hypothesis proposed by Toomre (1977), whereby major mergers of *gas-rich* disk galaxies represent the dominant process for producing the supermassive black hole and spheroid populations in the universe. Then, by the Soltan argument and the association of starbursts with quasars, it follows that this must also be the primary mechanism for triggering the most intense infrared luminous galaxies and the brightest quasars and active galactic nuclei (AGNs). It is important to keep in mind that this does not rule out other processes occurring at lower levels and under other circumstances. For example, we are not claiming that all AGNs result from mergers. In fact, low levels of such activity, as in Seyfert galaxies, often appear in undisturbed galaxies. For these objects, other modes of fueling are likely more significant, as in the stochastic accretion scenario of Hopkins & Hernquist (2006). The primary requirement on our model is that the bulk of the supermassive black hole mass density should have accumulated through

gas-rich mergers, consistent with the redshift evolution of the quasar population (Hopkins et al. 2007e). Similarly, spheroid evolution by gas-free ("dry") mergers will go on but does not explain how stellar mass is initially moved onto the red sequence or how black hole mass is initially accreted.

1.2. *Outline*

To test our hypothesis, we have developed methods for following the growth of black holes in numerical simulations of galaxy mergers, using a multiphase model for the star-forming gas that enables us to consider progenitor disks with large gas fractions. Generically, we find that major mergers of gas-rich galaxies evolve through distinct phases that can plausibly be identified with the various observed phenomena summarized above.

Figure 1 presents a schematic outline of these phases. In this picture, galactic disks grow mainly in quiescence, with the possibility of secular-driven bar or pseudobulge formation, until the onset of a major merger. A significant, perhaps even dominant, fraction of Seyferts and low-luminosity quasars will almost certainly arise from this secular evolution, but the prevalence of pseudobulges only in the hosts of ≲$10^7\ M_\odot$ black holes suggests this is limited to luminosities $M_B \gtrsim -23$ (see the discussion in § 3.3).

During the early stages of the merger, tidal torques excite some enhanced star formation and black hole accretion, but the effect is relatively weak, and the combination of large galactic dust columns and relatively small nuclear black holes means that only in rare circumstances (involving particular initial orbits and/or bulge-to-disk ratios) will the pair be identified as Seyferts or quasars. Most observationally identified mergers (and essentially all merging pairs) will be in this stage, and numerical simulations suggest it is the last stage at which the distinct nuclei enable automated morphological selection criteria to efficiently identify the system as a merger (Lotz et al. 2004; J. M. Lotz et al. 2008, in preparation). Care must therefore be taken with conclusions regarding the prevalence of starbursts and AGNs in these samples, as the small observed incidence of quasar activity (Dasyra et al. 2006; Myers et al. 2007b; Straughn et al. 2006; Alonso et al. 2007) is actually expected.

During the final coalescence of the galaxies, massive inflows of gas trigger starbursts with strengths similar to those inferred for ULIRGs and SMGs, although the actual mass in stars formed in these bursts is generally small compared to the stellar mass contributed by the merging disks. The high gas densities feed rapid black hole growth, but the black holes are obscured at optical wavelengths by gas and dust and are initially small compared to the newly forming spheroid. However, by the final stages, high accretion rate, heavily obscured (and in some cases nearly Compton-thick) BH growth in a ULIRG stage (often with merging binary BHs) appears ubiquitous (Komossa et al. 2003; Alexander et al. 2005a; Borys et al. 2005; Brand et al. 2006) and by high redshifts ($z \sim 2$) may dominate the obscured luminous quasar population (Alexander et al. 2005b; Stevens et al. 2005; Martínez-Sansigre et al. 2006; Brand et al. 2007).

Most of the nuclear gas is consumed by the starburst, and eventually feedback from supernovae and the black hole begins to disperse the residual gas. This brief transition or "blowout" phase will be particularly associated with highly dust-reddened (as opposed to more highly obscured type II) and/or IR-luminous quasars. As a relatively short phase, such objects constitute only ~20%–40% of the quasar population, similar to that observed (Gregg et al. 2002; White et al. 2003; Richards et al. 2003, 2006a; Hopkins et al. 2004). In fact, observational studies find that red
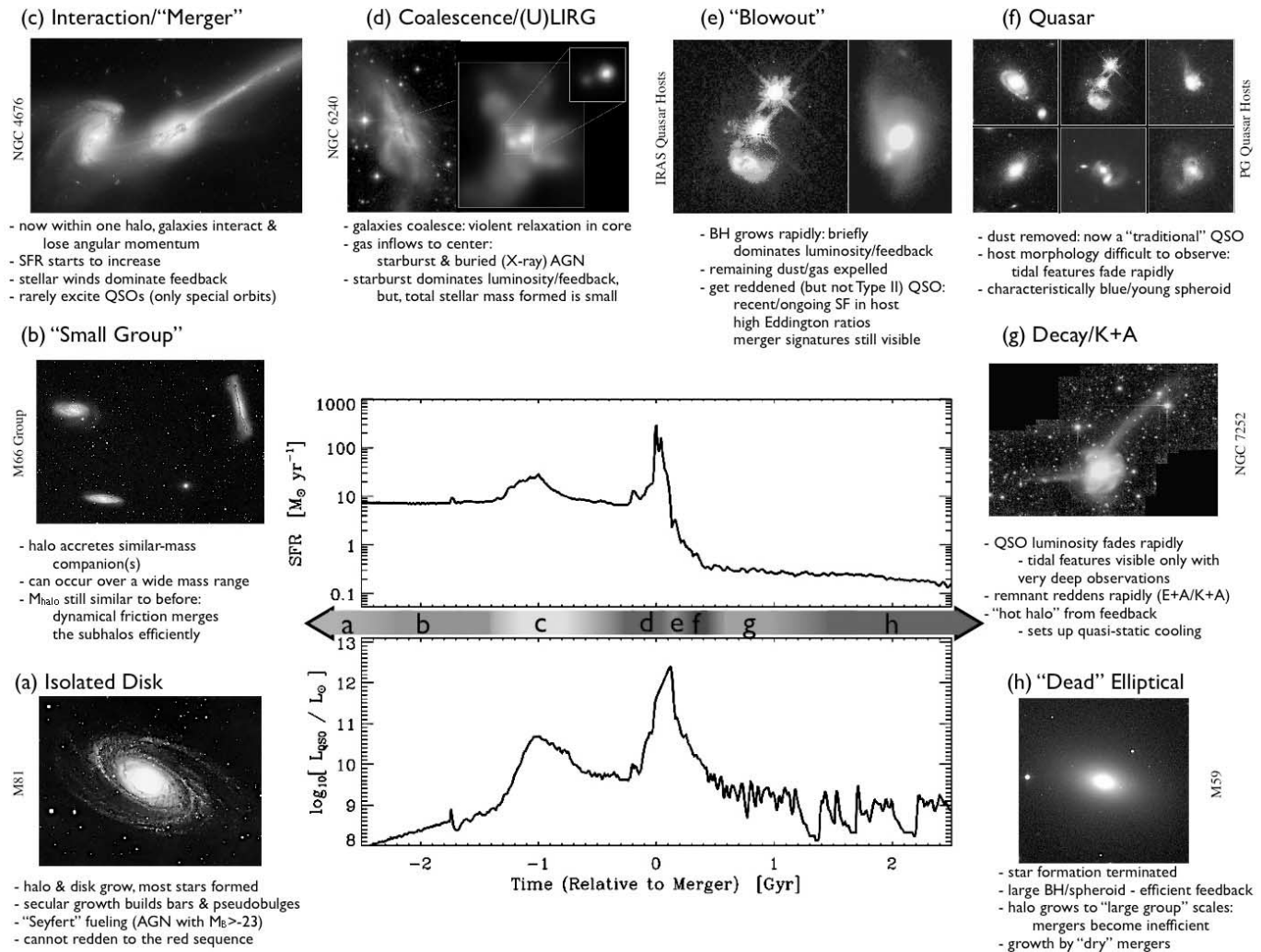
Fig. 1.—Schematic outline of the phases of growth in a "typical" galaxy undergoing a gas-rich major merger. Image credit: (*a*) NOAO/AURA/NSF; (*b*) REU program/NOAO/AURA/NSF; (*c*) NASA/STScI/ACS Science Team; (*d*) optical (*left*): NASA/STScI/R. P. van der Marel & J. Gerssen; X-ray (*right*): NASA/CXC/MPE/S. Komossa et al.; (*e*) *left*: J. Bahcall/M. Disney/NASA; *right*: Gemini Observatory/NSF/University of Hawaii Institute for Astronomy; (*f*) J. Bahcall/M. Disney/ NASA; (*g*) F. Schweizer (CIW/DTM); (*h*) NOAO/AURA/NSF. [*See the electronic edition of the Supplement for a color version of this figure.*]

quasar populations are related to mergers, with $\gtrsim 75\%$ (and as high as 100%) showing clear evidence of recent/ongoing merging (Hutchings et al. 2003, 2006; Kawakatu et al. 2006; Guyon et al. 2006; Urrutia et al. 2008), with young poststarburst stellar populations (Guyon et al. 2006), much of the dust arising on scales of the galaxy (in turbulent motions, inflow, and outflow; Urrutia et al. 2008), and extremely high Eddington ratios indicative of a still active period—making them (as opposed to most fully obscured quasars) a substantial contributor to the most luminous quasars in the universe (White et al. 2003; Hutchings et al. 2006; Zakamska et al. 2006). As the dust is removed, the black hole is then visible as a traditional optical quasar (although very small-scale "torus" obscuring structures may remain intact, allowing for some rare, bright type II systems).

Here observations of the host morphology are more ambiguous (see, e.g., Bahcall et al. 1997; Canalizo & Stockton 2001; Floyd et al. 2004; Zakamska et al. 2006; Pierce et al. 2007), but this is expected, for two reasons. First, the point-spread function of the bright and unobscured optical quasar must be subtracted and host galaxy structure recovered, a difficult procedure. Second, by this time the merger is complete and the spheroid has formed, leaving only fading tidal tails as evidence for the recent merger. Mock

observations constructed from the simulations (E. Krause et al. 2008, in preparation) imply that, with the best presently attainable data, these features are difficult to observe even locally and (for now) nearly impossible to identify at the redshifts of greatest interest ($z \gtrsim 1$). This appears to be borne out, as Bennert et al. (2008) have re-examined low-redshift quasars previously recognized from deep *Hubble Space Telescope* (*HST*) imaging as having relaxed spheroid hosts and found (after considerably deeper integrations) that every such object shows clear evidence for a recent merger. These difficulties will lead us to consider a number of less direct but more robust tests of the possible association between mergers and quasars.

Finally, as the remnant relaxes, star formation and quasar activity decline as the gas is consumed and dispersed, and the remaining galaxy resembles an elliptical with a quiescent black hole satisfying observed correlations between black hole and spheroid properties. During this intermediate $\sim$Gyr decay, depending on details of the merger and exact viewing time, the remnant may be classified as a low-luminosity (decaying) AGN in a massive (and relatively young) spheroid, or as a poststarburst (E+A/K+A) galaxy. Observationally, the link between K+A galaxies and mergers is well established (e.g., Yang et al. 2004; Goto 2005; Hogg et al.

2006 and references therein), and there is a clear tendency for these galaxies to host low-luminosity AGNs or LINERs (Yang et al. 2006; Goto 2006). Again, for the reasons given above, the situation is less clear for all low-luminosity AGNs (and there will be, as noted above, many such sources driven by secular mechanisms in disks). But more importantly most objects seen in this stage are expected to have relaxed to resemble normal spheroids. The merger exhausts gas and star formation in an immediate sense very efficiently, so the remnant reddens rapidly onto the red sequence. If this is also associated with quenching of future star formation (see Hopkins et al. 2007b, hereafter Paper II), then the spheroid will evolve passively, growing largely by dry mergers.

Individual simulations of mergers have enabled us to quantify the duration of these stages of evolution and how this depends on properties of the merging galaxies, such as their masses and gas content and the mass ratio and orbit of the encounter. In particular, we used the results to suggest a physical interpretation of quasar lifetimes (Hopkins et al. 2005d), to examine how quasars (Hopkins et al. 2005a) and starbursts (Chakrabarti et al. 2007) would evolve in this scenario, and to quantify structural properties of the remnant and how they depend on, e.g., the gas fractions of the merging galaxies (Cox et al. 2006a, 2006b; Robertson et al. 2006b, 2006c; Hopkins et al. 2007c).

In addition to making predictions for individual systems, we would also like to characterize how entire *populations* of objects would evolve cosmologically in our picture to test the model against the large body of observational data that exists from surveys of galaxies, quasars, and starbursts. Previously, we have adopted a semiempirical approach to this problem as follows. In our simulations, we can label the outcome by the final black hole mass in the remnant, $M_{BH,f}$, or, equivalently, the peak bolometric luminosity of the quasar, L_{peak}. Our simulations predict a regular behavior for the evolution of the different merger phases as a function of $M_{BH,f}$ or L_{peak} and also for the properties of the remnant as a function of $M_{BH,f}$ or L_{peak}. If we have an estimate of the observed distribution of systems in one phase of the evolution, we can then use our models to deconvolve the observations to infer the implied birthrate of such objects as a function of $M_{BH,f}$ or L_{peak}. Given this, the time behavior of the simulations provides a mapping between the different phases enabling us to make independent predictions for other populations. For example, knowing the observed quasar luminosity function (QLF) at some redshift, our simulations allow us to predict how many quasar-producing mergers of a given mass must be occurring at the time, which can then be tested against the observed merger statistics.

We exploited this approach to examine the relationship between the abundance of quasars and other manifestations of quasar activity and showed that our model for quasar lifetimes and light curves yields a means to interpret the shape of the QLF (Hopkins et al. 2005b), provides a consistent explanation for observations of the QLF at optical and X-ray frequencies (Hopkins et al. 2005c), explains observed evolution in the faint-end slope of the QLF (Hopkins et al. 2006b), and can account for the spectral shape of the cosmic X-ray background (Hopkins et al. 2006a, 2007e). Using this technique to map between different types of objects, we demonstrated that the observed evolution and clustering of the quasar population is consistent with observations of red galaxies (Hopkins et al. 2006c, 2006d, 2007d) and merging systems (Hopkins et al. 2006f, 2007a), as well as the mass function of supermassive black holes and its estimated evolution with redshift (Hopkins et al. 2006a, 2007e). In each case, we found good agreement with observations provided that the mappings were based on the lifetimes and light curves from our merger

simulations and not idealized ones that have typically been used in earlier theoretical studies. We further showed that our picture makes numerous predictions (Hopkins et al. 2006a, 2007a) that can be used to test our hypothesis, such as the luminosity dependence of quasar clustering (Lidz et al. 2006). However, the cosmological context of our results was not provided in an entirely theoretical manner because our analysis relied on an empirical estimate of one of the connected populations.

Obtaining a purely theoretical framework for our scenario is difficult because cosmological simulations including gas dynamics currently lack the resolution to describe the small-scale physics associated with disk formation, galaxy mergers, star formation, and black hole growth. Semianalytic methods avoid some of these limitations, but at the expense of parameterizing the unresolved physics in a manner this is difficult to calibrate independently of observational constraints. For the time being, neither approach is capable of making an entirely ab initio prediction for how the various populations we are attempting to model would evolve with time.

In this paper, we describe a strategy that enables us, for the first time, to provide a purely theoretical framework for our picture. Our procedure is motivated by, but does not rely on, observations suggesting that there is a characteristic halo mass hosting bright quasars. This inference follows from measurements of the clustering of quasars in the 2dF, SDSS, and other surveys (Porciani et al. 2004; Porciani & Norberg 2006; Wake et al. 2004; Croom et al. 2005; Coil et al. 2007; Myers et al. 2007a; da Angela et al. 2007; Shen et al. 2007) and investigations of the quasar proximity effect (Faucher-Giguere et al. 2008; Kim & Croft 2007; Nascimento Guimaraes et al. 2007). By adopting simple models for the merger efficiency of galaxies as a function of environment and mass ratio, we show that this characteristic halo mass for quasars corresponds to the most favorable environment for major mergers between gas-rich disks to occur, namely the "small group" scale. This finding argues for an intimate link between such mergers and the triggering of quasar activity and naturally leads to a method for determining the redshift evolution of the quasar population from dark matter simulations of structure formation in a ΛCDM universe.

By combining previous estimates of the evolution of the halo mass function with halo occupation models and our estimates for merger timescales, we infer the statistics of mergers that excite quasar activity. We then graft onto this our modeling of quasar light curves and lifetimes, obtained from our simulations of galaxy mergers that include star formation and black hole growth to deduce, in an ab initio manner, the redshift-dependent birthrate of quasars as a function of their peak luminosities and the corresponding formation rate of black holes as a function of mass. Because our merger simulations relate starbursts, quasars, and red galaxies as different phases of the same events, we can then determine the cosmological formation rate of these various populations and their evolution with redshift. In particular, as we demonstrate in what follows, the observed abundance of all these objects is well matched to our estimates, unlike for other theoretical models, supporting our interpretation that mergers between gas-rich galaxies represent the dominant production mechanism for quasars, intense starbursts, supermassive black holes, and elliptical galaxies.

We investigate this in a pair of companion papers. Here (Paper I), we describe our model and use it to investigate the properties of mergers and merger-driven quasar activity. In the companion paper (Paper II), we extend our study to the properties of merger remnants and the formation of the early-type galaxy population. Specifically, § 2 outlines our methodology, describing the physical

criteria for and identification of major mergers (§ 2.1), the distribution of mergers across different scales and galaxy types (§ 2.2), and the dependence of mergers on environmental properties (§ 2.3). We then examine the predicted merger mass functions, fractions, and clustering properties from this model, and we compare with observations to verify that we are appropriately modeling the merger history of the universe (§ 2.4). In § 3 we examine the consequences of a general model in which mergers trigger quasar activity. We present a number of robust predictions both independent of (§ 3.1) and including (§ 3.2) physical models for the quasar light curves and duty cycles in mergers. We contrast this with a "secular" model in which quasar activity is caused by disk instabilities (§ 3.3) and show that a variety of independent constraints suggest that such a mode cannot dominate the formation of bright, high-redshift quasars. We discuss and summarize our conclusions in § 4.

Throughout, we adopt a WMAP3 $(\Omega_M, \Omega_\Lambda, h, \sigma_8, n_s) = (0.268, 0.732, 0.704, 0.776, 0.947)$ cosmology (Spergel et al. 2007) and normalize all observations and models shown to these parameters. Although the exact choice of cosmology may systematically shift the inferred bias and halo masses (primarily scaling with σ_8), our comparisons (i.e., relative biases) are for the most part unchanged, and repeating our calculations for a "concordance" $(0.3, 0.7, 0.7, 0.9, 1.0)$ cosmology or the WMAP1 $(0.27, 0.73, 0.71, 0.84, 0.96)$ results of Spergel et al. (2003) has little effect on our conclusions. We also adopt a diet Salpeter IMF following Bell et al. (2003b) and convert all stellar masses and mass-to-light ratios accordingly. Again, the choice of the IMF systematically shifts the normalization of stellar masses herein but does not substantially change our comparisons. *UBV* magnitudes are in the Vega system, and SDSS *ugriz* magnitudes are AB.

2. MERGERS

2.1. *What Determines whether Galaxies Merge*

2.1.1. *Physical Processes*

To begin, we postulate which mergers are relevant to our picture. Minor mergers (mass ratios ≫3:1) will not trigger significant star formation or quasar activity for most orbits, and consequently will neither exhaust a large fraction of the larger galaxy's gas supply nor be typically identified as mergers observationally. We are therefore specifically interested in major mergers, with mass ratios ≤3:1, but note that our conclusions are unchanged if, instead of this simple threshold, we include all mergers and adopt some mass-ratio dependent efficiency (e.g., assuming the fractional BH/bulge growth scales with mass ratio R in some power-law fashion, $\propto R^{-1}$, as suggested by numerical simulations; J. Younger et al. 2008, in preparation). In this case, the decreasing efficiency of BH fueling in minor mergers leads (as expected) to the conclusion that they are only important at low masses/luminosities (similar to where secular activity may dominate quasar populations; see § 3.3), and our predictions for massive bulges and BHs are largely unaffected. If the timescale for two galaxies to merge is long compared to the Hubble time, they clearly will not have merged in the actual universe. However, the merger timescale must also be short compared to the time required to tidally strip or disrupt either of the galaxies—if it is not, then by the time the galaxies finally coalesce, the end result will simply be tidal accretion of material at large radii.

This defines two fundamental criteria for galaxy mergers to occur in the setting of a halo of mass M_{halo}:

1. The halo must host at least two galaxies of comparable mass $\sim M_{gal}$. Note that even for mergers of distinct host halos in the field, the halo-halo merger proceeds much faster than the merger of the galaxies, so there is some period where the two can be considered distinct substructures or distinct galaxies within a common host.

2. The merger must be efficient—i.e., occur in much less than a Hubble time. This requires that the mass of the galaxies and their associated (bound) dark matter subhalos be comparable to the mass of the parent halo (e.g., for the simplest dynamical friction arguments, requiring $M_{halo}/M_{gal} \ll 30$).

Together, these criteria naturally define a preferred mass scale for major mergers (host halo mass M_{halo}) for galaxies of mass M_{gal}. A halo of mass $\langle M_{halo}\rangle(M_{gal})$ typically hosts a galaxy of mass M_{gal}. At smaller (relative) halo masses $M_{halo} \ll \langle M_{halo}\rangle$, the probability that the halo hosts a galaxy as large as M_{gal} declines rapidly (and eventually must be zero or else violate limits from the cosmic baryon fraction). At larger $M_{halo} \gg \langle M_{halo}\rangle$, the probability that the halo will merge with or accrete another halo hosting a comparable $\sim M_{gal}$ galaxy increases, but the efficiency of the merger of these galaxies declines rapidly. Eventually, the M_{gal} galaxies are relatively small satellites in a large parent halo of mass $M_{halo} \gg \langle M_{halo}\rangle$, for which (satellite-satellite) mergers are extremely inefficient (given the high virial velocities of the host, and dynamical friction timescales $\gg t_H$).

The preferred major-merger scale for galaxies of mass M_{gal} is therefore only slightly larger (factor ~ 2) than the average host halo mass for galaxies of this mass. We refer to this as the small group scale and emphasize the term *small* in this name: the average halo of this mass still hosts only one galaxy of mass $\sim M_{gal}$, and the identifiable groups will only consist of 2–3 members of similar mass (although there may of course be several much smaller systems in the group, which have little dynamical effect). This is very different from large group scales, easily identified observationally, which consist of ≫3 members.

Figure 2 illustrates several of these points. We adopt the merger timescales derived below and use the halo occupation fits from Wang et al. (2006) to determine the probability of a halo hosting a pair of galaxies of a given mass: the details of the formalism are described below and used throughout, but we wish to illustrate the key qualitative points. The merger timescale for galaxies of a given mass is shortest when they are large relative to their host halo mass, as expected from dynamical friction considerations. However, the probability of a pair being hosted cuts off sharply at low halo masses. Moreover, the contribution to mergers of galaxies of mass M_{gal} from larger halos is further suppressed by the simple fact that there are fewer halos of larger masses.

Modern, high-resolution dark matter–only cosmological simulations (e.g., Springel et al. 2005c) have made it possible to track the merger histories of galaxy halos over large ranges in cosmic time and halo mass. For our purposes, the critical information is contained in the subhalo mass function, which has been quantified in great detail directly from such simulations (Kravtsov et al. 2004; Gao et al. 2004; Nurmi et al. 2008) and from extended Press-Schechter theory and semianalytic approaches (Taylor & Babul 2004; Zentner et al. 2005; van den Bosch et al. 2005) calibrated against numerical simulations.

When a halo (containing a galaxy and its own subhalo populations) is accreted, the accretion process is relatively rapid—the accreted halo will always be identifiable for *some* period of time as a substructure in the larger halo. Although the new subhalo may lose mass to tidal stripping, there will still be some dark matter subhalo associated with the accreted galaxy, which will remain until the substructure merges with the central galaxy via dynamical friction or (much more rarely) another satellite substructure.
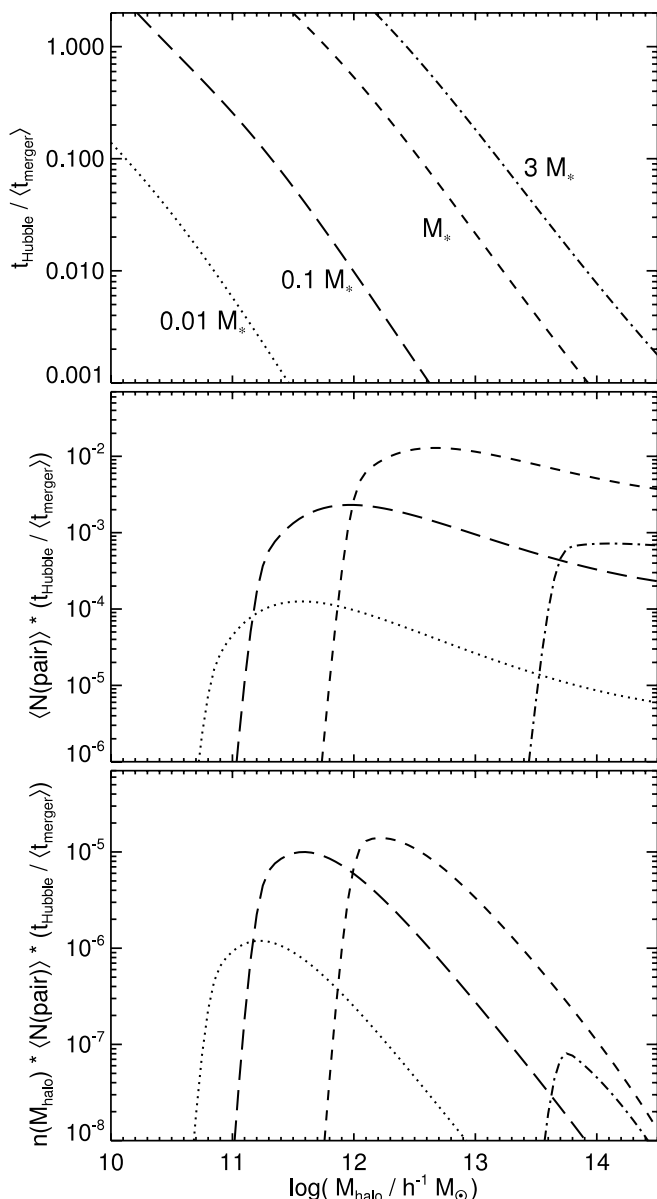
Fig. 2.—Efficiency of major galaxy mergers (of a certain galaxy mass relative to the characteristic local Schechter function M_*) as a function of host halo mass (at $z = 0$, but the results are qualitatively similar at all redshifts). *Top*: Merger timescale relative to the Hubble time (assuming a pair of galaxies of mass M_{gal} are hosted in a halo of mass M_{halo})—mergers occur rapidly ($t_{merger} \ll t_H$) when the halo mass is small relative to the galaxy mass (we temporarily ignore the obvious requirement that $M_{gal} < f_{baryon} M_{halo}$). *Middle*: Same, but now multiplied by the probability that the halo actually hosts a pair of galaxies of the given mass (technically, within a mass ratio 3:1), given the empirical halo occupation model from Wang et al. (2006). Although mergers are most rapid in the lowest-mass halos, these halos do not host relatively massive galaxies. *Bottom*: Same, but further multiplied by the abundance of halos of a given mass—the fact that the halo mass function and merger efficiency are decreasing functions of M_{halo} (for fixed M_{gal}) means that the contribution to galaxy mergers of a given M_{gal} will be dominated by the lowest-mass halos in which there is a significant probability to accrete/host a pair of M_{gal} galaxies—the small group scale.

Therefore, knowing the subhalo populations of all halos at a given instant, the calculation of the rate and distribution of *galaxy* mergers depends only on calculating the efficiency of the subhalo/galaxy mergers within these halos. This is a great advantage—we do not need to calculate halo-halo merger rates, which are not well defined (even when extracted directly from cosmological simulations) and depend sensitively on a number of definitions (see, e.g., Gottlöber et al. 2001; Maller et al. 2006), but instead we work from the robust

(and well-defined) subhalo mass function (see Gao et al. 2004 and references therein).

This is similar to many of the most recent semianalytic models, which adopt a hybrid approach to determine galaxy mergers, in which galaxies survive independently so long as their host halo remains a distinct substructure, after which point a dynamical friction "clock" is started and the galaxy merges with the central galaxy in its parent halo at the end of the dynamical friction time. Fortunately, for our purposes we are only interested in major mergers with mass ratios $\lesssim 3:1$. In these cases, dynamical friction acts quickly on the subhalos (infall time $\lesssim t_H/3$ at all redshifts), and the primary ambiguity will be the *galaxy* merger time in their merged or merging subhalos.

To perform this calculation, we need to know the properties of the merging galaxies. For now, we only want to calculate where and when galaxies are merging, not how they evolved to their present state in the first place. This is our primary reason for not constructing a full semianalytic model: rather than introduce a large number of uncertainties, theoretical prescriptions that we are not attempting to test here, and tunable parameters in order to predict that, e.g., a $10^{11} M_\odot$ halo typically hosts a $\sim 10^{10} M_\odot$ star-forming galaxy, we can adopt the established empirical fact that this is so. In detail, we populate subhalos according to an empirical halo occupation model (e.g., Tinker et al. 2005; Conroy et al. 2006; Vale & Ostriker 2006; van den Bosch et al. 2007; Wang et al. 2006)—i.e., matching the observed statistics of where galaxies of a given type live (accounting for different occupations for different galaxy types/colors, and the scatter in galaxies hosted in halos of a given mass).

This is sufficient for most of our predictions. We do not necessarily need to know exactly how long it will take for these mergers to occur, only that they are occurring at a given redshift—i.e., that the objects will merge and that the merger time is shorter than the Hubble time (which for the mass ratios of interest is essentially guaranteed). For example, predicting the clustering of galaxy mergers does not require knowledge of how rapidly they occur, only *where* they occur. Even predicting the observed merger mass function does not rely sensitively on this information, since the duration over which the merger is visible will be comparable (albeit not exactly equal) to the duration over which the merger occurs (such that a fixed fraction ~ 1 of all merging systems are observable).

However, for the cases where it is necessary, we estimate the timescales for the galaxies to merge and to be identified as mergers. This is the most uncertain element in our model. Part of this uncertainty is due to the large parameter space of mergers (e.g., differences in orbital parameters, relative inclinations, etc.). These uncertainties are fundamental but can at least be controlled by comparison to large suites of hydrodynamic simulations that sample these parameter spaces (Robertson et al. 2006b) and allow us to quantify the expected range of merger properties due to these (essentially random) differences. The more difficult question is how appropriate any analytic merger timescale or cross section can be. To address this, we will throughout this paper consider a few representative models:

Dynamical friction.—The simplest approximation is that the galaxies are point masses, and (once their subhalos merge) they fall together on the dynamical friction timescale. This is what is adopted in most semianalytic models. In fact, this is only an appropriate description when the galaxies are small relative to the enclosed halo mass and are both moving to the center of the potential well—which is often not the case at these late stages. While unlikely to be incorrect by orders of magnitude, this approximation begins to break down when the galaxies are relatively large

compared to their halos (common in $\lesssim 10^{12}\ M_\odot$ halos) and when the galaxies are very close (and could, e.g., enter a stable orbit). What finally causes galaxies to merge is not, in fact, simple dynamical friction but dissipation of angular momentum via a resonance between the internal and orbital frequencies.

Group capture (collisional).—On small scales, in satellite-satellite mergers, or in the merger of two small field halos, it is more appropriate to consider galaxy mergers as a collisional process in which there is some effective gravitational cross section. In other words, galaxy mergers proceed once the galaxies pass at sufficiently small distances with sufficiently low relative velocity. There have been a number of theoretical estimates of these cross sections—we adopt here the fitting formulae from Krivitsky & Kontorovich (1997), who calibrate the appropriate cross sections from a set of numerical simulations of different encounters and group environments. This compares well with other calculations (White 1976; Makino & Hut 1997; Mamon 2006 and references therein), and we find little difference using these alternative estimations. For large mass ratios and separations, the expressions appropriately reduce to the dynamical friction case.

Angular momentum.—Binney & Tremaine (1987) consider this problem from the perspective of the angular momentum-space in which galaxy mergers are allowed. This approach is similar to the capture estimates above but accounting for capture into orbits as well. Whether or not such orbits will merge is, of course, somewhat ambiguous—it is likely that some significant fraction are stable and will not merge, while others decay rapidly due to resonance between the disk circular frequencies and the orbital frequency. Nevertheless, this serves to bracket the range of likely merger configurations.

2.1.2. Synopsis of Model and Uncertainties

Thus, to summarize our approach: at a given redshift, we calculate the halo mass function $n(M_{halo})$ for our adopted cosmology following Sheth et al. (2001). For each halo, we calculate the (weakly mass- and redshift-dependent) subhalo mass function (or distribution of subhalos, $P[N_{subhalo}|M_{subhalo}, M_{halo}]$) following Zentner et al. (2005) and Kravtsov et al. (2004). Alternatively, we have adopted it directly from Gao et al. (2004) and Nurmi et al. (2008) or calculated it following van den Bosch et al. (2005) and Vale & Ostriker (2006) and obtain similar results. Note that the subhalo masses are defined as the masses on accretion by the parent halo, which makes them a good proxy for the hosted galaxy mass (Conroy et al. 2006) and removes the uncertainties due to tidal mass stripping.

Mergers are identified by the basic criteria described above. We populate these halos and subhalos with galaxies following the empirical halo occupation models of Conroy et al. (2006; see also Vale & Ostriker 2006) normalized directly with group observations following Wang et al. (2006) at $z = 0$ (considering instead the occupation fits in Yang et al. [2003], Cooray [2005, 2006], Zheng et al. [2005], and van den Bosch et al. [2007] makes little difference). This determines both the mean stellar mass and dispersion in stellar masses of galaxies hosted by a given halo/subhalo mass $P(M_{gal}|M_{subhalo})$, which (optionally) can be broken down separately for blue and red galaxy types.

Figure 3 shows the mean galaxy mass as a function of halo mass from this model at $z = 0$. Since the halo occupation models consider stellar mass or luminosity, we use the baryonic and stellar mass Tully-Fisher relations calibrated by Bell & de Jong (2001) to convert between the two. (We have also compared the global baryonic mass function estimated in this manner with that observationally inferred in Bell et al. [2003a] and find good agreement).
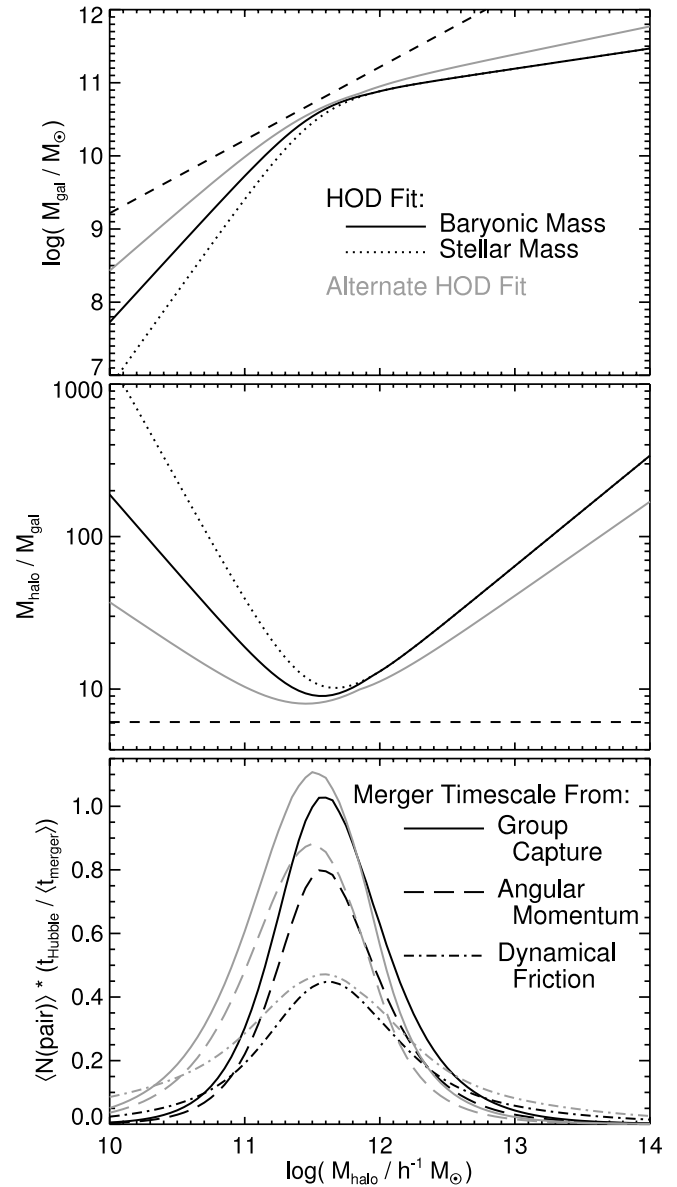

Fig. 3.—Illustration of basic elements of importance to where galaxy-galaxy mergers occur. *Top*: Average central galaxy stellar (*dotted line*) and baryonic (*solid line*) mass as a function of host halo mass, in our typically adopted halo occupation model (Conroy et al. 2006, Vale & Ostriker 2006, *black*), and the alternate halo occupation model from Yang et al. (2003, *gray*); only baryonic mass shown. *Middle*: Corresponding halo-to-galaxy mass ratio. *Bottom*: Average major merger timescale/efficiency [calculated as in the middle panel of Fig. 2, but for the appropriate mean $M_{gal}(M_{halo})$]. Timescales are determined as described in the text, from dynamical friction (*dot-dashed line*), group capture (*solid line*), or angular momentum (*long-dashed line*) considerations. [*See the electronic edition of the Supplement for a color version of this figure.*]

If necessary, we calculate the galaxy-galaxy merger efficiency/timescale using the different estimators described above. Figure 3 also shows the expected merger efficiency as a function of halo mass for these mean values (i.e., the probability of hosting a subhalo within the appropriate mass range convolved with the calculated merger timescale). The qualitative features are as expected from Figure 2. The different merger timescale estimators agree well at large halo masses, with the dynamical friction treatment yielding a somewhat longer (factor $\lesssim$ few) timescale at intermediate masses (but this is near the regime of low M_{halo}/M_{gal}, where the dynamical friction approximation is least accurate).

The main elements and their uncertainties in our model are as follows:

1. *Halo mass function.*—We begin by computing the overall halo mass function. There is very little ambiguity in this calculation at all redshifts and masses of interest ($z \lesssim 6$; see, e.g., Reed et al. 2007), and we do not consider it a significant source of uncertainty.

2. *Subhalo mass function.*—The subhalo mass function of each halo is then calculated. Although numerical simulations and semianalytic calculations generally give very similar results (especially for the major-merger mass ratios of interest in this paper, as opposed to very small subhalo populations; see van den Bosch et al. 2005), there is still some (typical factor <2) disagreement between different estimates. We therefore repeat most of our calculations adopting both our "default" subhalo mass function calculation (Zentner et al. 2005; Kravtsov et al. 2004) and an alternative subhalo mass function calculation (van den Bosch et al. 2005, normalized to match cosmological simulations as in Shaw et al. 2006), which brackets the range of a number of different estimates (e.g., Springel et al. 2001; Tormen et al. 2004; De Lucia et al. 2004; Gao et al. 2004; Nurmi et al. 2008), and demonstrate the uncertainty due to this choice. The difference is ultimately negligible at $M_{gal} \gtrsim 10^{10}\ M_{\odot}$ at all redshifts and rises to only a factor of ~ 2 at $M_{gal} \lesssim 10^{10}\ M_{\odot}$ (probably owing to differences in the numerical resolution of different estimates at low halo masses).

3. *Halo occupation model.*—We then populate the central galaxies and "major" subhalos with an empirical halo occupation model. Although such models are constrained, by definition, to reproduce the mean properties of the halos occupied by galaxies of a given mass/luminosity, there are known degeneracies between parameterizations that give rise to (typical factor of ~ 2) differences between models. We therefore again repeat all our calculations for our "default" model (Conroy et al. 2006; see also Vale & Ostriker 2006) and an alternate halo occupation model (Yang et al. 2003; see also Yan et al. 2003; Zheng et al. 2005), which bracket the range of a number of calculations (e.g., Cooray 2005, 2006; Zheng et al. 2005; van den Bosch et al. 2007). Again, we find this yields negligible differences at $M_{gal} \gtrsim 10^{10}\ M_{\odot}$ (as the clustering and abundances of massive galaxies are reasonably well constrained, and most of these galaxies are central halo galaxies), and even at low masses the typical discrepancy rises to only ~ 0.2 dex.

We note that we have also considered a variety of prescriptions for the redshift evolution of the halo occupation model: including that directly prescribed by the quoted models, a complete rederivation of the HOD models of Conroy et al. (2006) and Vale & Ostriker (2006) at different redshifts from the observed mass functions of Fontana et al. (2006), Bundy et al. (2005), Borch et al. (2006), and Blanton (2006) (see § 3.1), or simply assuming no evolution (in terms of galaxy mass distributions at fixed halo mass, for either all galaxies or star-forming galaxies). We find that the resulting differences are small (at least at $z \lesssim 3$), comparable to those inherent in the choice of halo occupation model. This is not surprising, as a number of recent studies suggest that there is very little evolution in halo occupation parameters (in terms of mass, or relative to L_*) with redshift (Yan et al. 2003; Cooray 2005; Conroy et al. 2006), or equivalently that the masses of galaxies hosted in a halo of a given mass are primarily a function of that halo mass and not of redshift (Heymans et al. 2006; Conroy et al. 2007). This appears to be especially true for star-forming and $\sim L_*$ galaxies (of greatest importance for our conclusions; Conroy et al. 2007), unsurprising given that "quenching" is not strongly operating in those systems to change their mass-to-light ratios.

4. *Merger timescale.*—Having populated a given halo and its subhalos with galaxies, we then calculate the timescale for mergers between major galaxy pairs. This is ultimately the largest source of uncertainty in our calculations, at all redshifts and masses. Again, we emphasize that some of our calculations are completely independent of these timescales. However, where adopted, we illustrate this uncertainty by presenting all of our predictions for three estimates of the merger timescale: a simple dynamical friction formula, a group capture or collisional cross section estimate, and an angular momentum (orbital cross section) capture estimate, all as described above. At large masses and redshifts $z \lesssim 2.5$, this is a surprisingly weak source of uncertainty, but the estimated merger rates/timescales can be very different at low masses $M_{gal} \lesssim 10^{10}\ M_{\odot}$ and the highest redshifts $z \sim 3-6$.

At low masses, this is due to to a variety of effects, including the substantial difference between infall or merger timescales and the timescale for morphological disturbances to be excited (different in, e.g., an impact approximation as opposed to the circular orbit decay assumed by dynamical friction).

The difference in redshift evolution is easily understood: at fixed mass ratio, the dynamical friction timescale scales as $t_{df} \propto t_H \propto \rho^{-1/2}$, but a "capture" timescale will scale with fixed cross section as $t \propto 1/(n\langle\sigma v\rangle) \propto \rho^{-1}$, so that (while the details of the cross sections and dependence of halo concentration on redshift make the difference not quite as extreme as this simple scaling) the very large densities at high redshift make collisional merging increase rapidly in efficiency. The true solution is probably some effective combination of these two estimates, and the "more appropriate" approximation depends largely on the initial orbital parameters of the subhalos. At present, we therefore must recognize this as an inherent uncertainty, but one that serves to bracket the likely range of possibilities at high redshifts.

2.2. *Where Mergers Occur*

We are now in a position to predict the statistics of mergers. First, we illustrate some important qualitative features. Figure 4 shows the merger efficiency (as in Fig. 2) for different classes of mergers: major mergers with the central galaxy in a halo, minor mergers with the central galaxy, and major mergers of two satellite galaxies in the halo. We show the results for our "default" model, adopting the dynamical friction merger timescale, but the qualitative results are independent of these choices. The key features are expected: major mergers are efficient at small group scales (halo masses) comparable to or just larger than the average host halo mass for a given M_{gal}. At larger M_{halo}, major mergers become more rare for the reasons in § 2.1. However, although dynamical friction times increase, the rapidly increasing number of satellite systems in massive halos means that minor merger accretion onto the central galaxy proceeds with a relatively constant efficiency. This will not trigger substantial quasar or starburst activity or morphological transformation, but it may be important for overall mass growth in large cD galaxies, although recent cosmological simulations (Maller et al. 2006) suggest that major mergers dominate minor mergers in the assembly of massive galaxies (although their simulation does not extend to the largest cD galaxies).

Satellite-satellite minor mergers are a small effect at all masses, as expected (by the time a halo is sufficiently massive to host a large number of satellites of a given M_{gal}, the orbital velocity of the galaxies about the halo is much larger than their individual
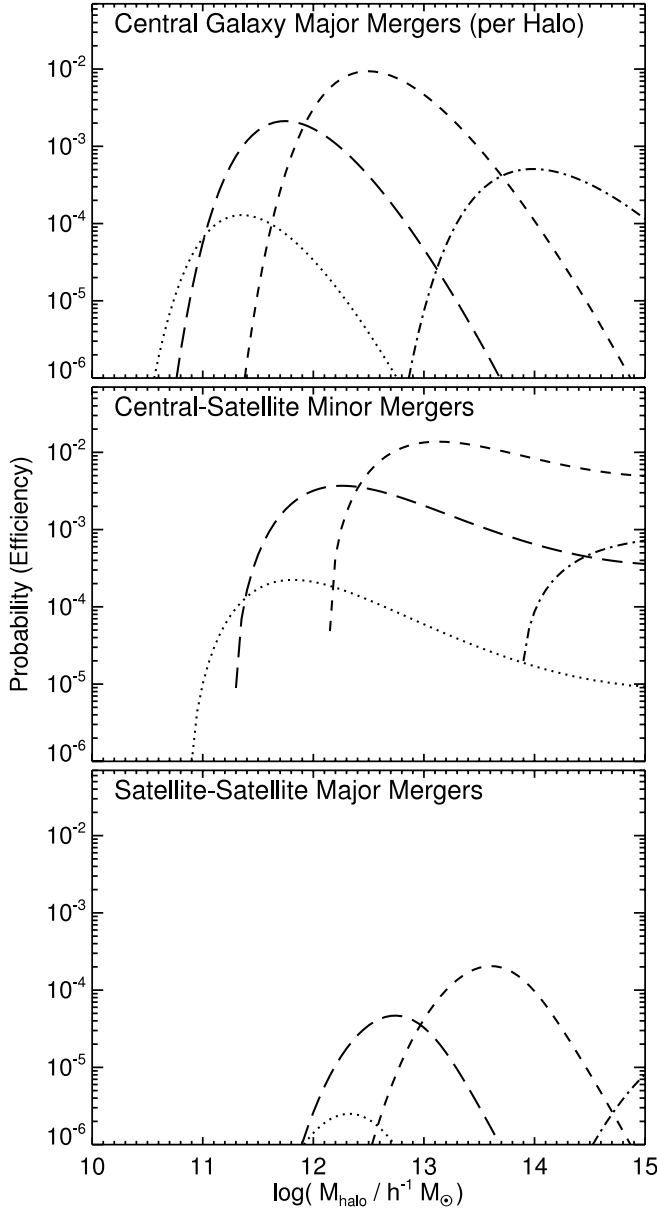
FIG. 4.—Merger efficiency (arbitrary units; defined in the same manner as the lower panel of Fig. 2, with different line styles in the same style for various mass galaxies) for different classes of mergers. Using the subhalo mass functions and halo occupation models, we can separate major mergers onto the central galaxy in a halo (*top*), minor (mass ratio $>3:1$ but $<10:1$) mergers onto the central galaxy (*middle*), and satellite-satellite mergers (*bottom*). Major mergers occur efficiently in central galaxies near the small group scale for each M_{gal}. When galaxies live in very massive halos, they experience a large number of minor mergers from the satellite population. Satellite-satellite mergers are a relatively small effect at all galaxy and halo masses.

FIG. 5.—Merger efficiency (arbitrary units; calculated as in Fig. 2) as a function of halo mass [adopting the mean $M_{gal}(M_{halo})$ from Fig. 3]. Using the type-separated galaxy mass functions from Bell et al. (2003b), Borch et al. (2006), and Fontana et al. (2004) at $z = 0,\ 1,\ 2$, respectively, we show the fraction of galaxies at each mass expected to be gas-rich and gas-poor, at each of three redshifts. At high redshifts, all but the most massive merging galaxies will be gas-rich, whereas at low masses the gas-poor population dominates at most masses where mergers are efficient. [*See the electronic edition of the Supplement for a color version of this figure.*]

internal velocities). In what follows, we will generally ignore satellite-satellite mergers. Including them is a very small correction (generally $\ll 10\%$), and their dynamics are uncertain. Moreover, their colors and star formation histories are probably affected by processes such as tidal stripping, harassment, and ram-pressure stripping, which we are neither attempting to model nor test. We have, however, checked that there are no significant or qualitative changes to our predictions if we (naively) include the satellite-satellite term.

Although the consequences of the merger will be very different, the efficiency with which two galaxies merge does not depend strongly on whether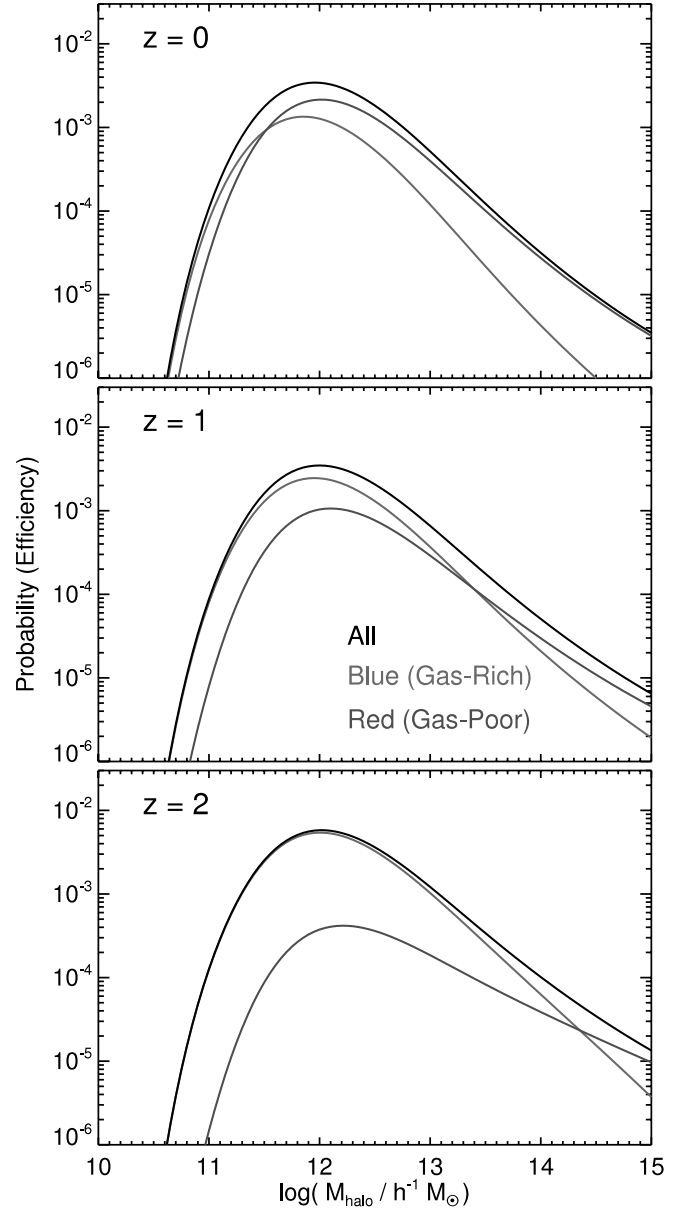 they are star-forming or red/passive (all else being equal). It is therefore a consequence that, at low redshifts, gas-rich mergers are generally relegated to low stellar masses and field environments where such galaxies are common. Figure 5 illustrates this. We plot the mean efficiency of major, central galaxy mergers (as in Fig. 4, but for the mean M_{gal} at each M_{halo}) as a function of halo mass at each of three redshifts. At each redshift, we divide this into the observed fraction of red and blue galaxies at the given galaxy/halo mass, using the appropriate observed, type-separated galaxy mass functions. The efficiency of mergers at a given halo and galaxy mass does not evolve (note that this is *not* a statement that the overall merger rates will not change, but rather a statement that the same galaxies in the same halos will merge at the same rate). However, at low redshifts, red galaxies
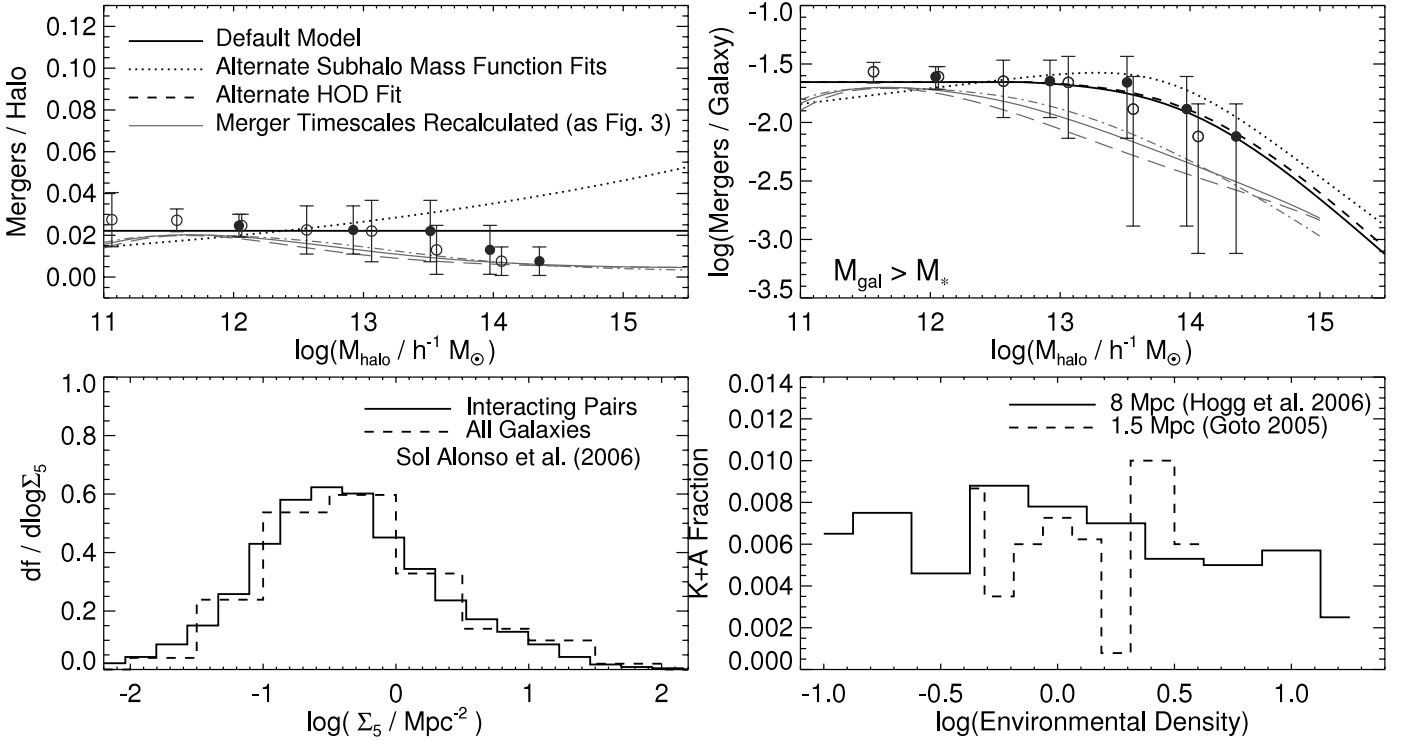
Fig. 6.—*Top*: Merger fraction as a function of host halo mass. The fraction of all halos (groups) predicted to host at least one major merger of galaxy mass $\gtrsim 10^{10} M_\odot$ is plotted (*left*), as is the fraction of all galaxies in halos of a given M_{halo} that are merging (*right*). We show the predictions for several variations of our standard model (described in the text) used to identify all merging systems (*black lines, as labeled*), and adding a more detailed calculation of the actual timescale for the physical galaxy mergers (*gray lines, as labeled*) and ability to morphologically identify them. Both are compared with observed merger fractions (*points*) from Sol Alonso et al. (*gray circles*, 2006) (we convert their measured intermediate-scale densities to average halo masses following Baldry et al. 2006 and Kauffmann et al. 2004 shown as open and filled points, respectively). *Bottom*: The observed distributions (fraction of objects per logarithmic interval in galaxy surface density) of merger and normal galaxy environments, from the group catalogs of Sol Alonso et al. (2006, *left*), and the fraction of recent merger remnant (poststarburst, K+A) galaxies as a function of galaxy surface density averaged on intermediate (1.5 Mpc) and large (8 Mpc) scales (*right*). Mergers occur on all scales and in halos of all masses, without a strong feature at a particular scale. [*See the electronic edition of the Supplement for a color version of this figure.*]

dominate the mass budget, whereas at high redshifts, most galaxies are still blue (star-forming) in all but the most massive halos. We will discuss the possibility that mergers themselves drive this change in the blue and red fractions in Paper II, but for now we illustrate that the locations of gas-rich and dry mergers reflect where gas-rich and gas-poor galaxies dominate the population, respectively, which is empirically determined at the redshifts of interest here. We note that our halo occupation models do not explicitly model a dependence of halo populations on central galaxy properties; i.e., the tentative observational suggestion that, at fixed halo and galaxy mass, red central galaxies are preferentially surrounded by red (as opposed to blue) satellites (Weinmann et al. 2006). If real, the effect of such a trend is to make the transition plotted in Figure 4 somewhat sharper—this has little effect on our conclusions but does somewhat lower the predicted gas-rich merger rates (and corresponding predicted quasar luminosity density) at $z \lesssim 0.5$ (since a red central galaxy would have a lower probability of an infalling, gas-rich system).

Integrating over the appropriate galaxy populations, Figure 6 compares the predicted $z = 0$ merger fraction as a function of halo mass from this model with that observed. The agreement is good over a wide dynamic range. Although there is a significant (factor of ~ 2) systematic difference based on how this fraction is calculated, this is within the range of present observational uncertainty. It is also important to distinguish the merger fraction of parent halos (i.e., fraction of groups that contain a merger) and that of galaxies (i.e., fraction of all galaxies at a given M_{gal} or M_{halo} that are merging), as at large halo masses the rate of mergers onto the central galaxy could remain constant (giving a constant merger

rate per halo), but the inefficient merging of the increasingly large number of satellites will cause the galaxy merger fraction to fall rapidly.

We also show the distribution of mergers (interacting pairs) and all galaxies in environmental density (local projected surface density $\Sigma_5 = 5/(\pi d_5^2)$, where d_5 is the distance to the fifth nearest neighbor) from the local group catalogs of Sol Alonso et al. (2006)—we compare this data set directly to our prediction by converting Σ_5 to M_{halo} using the mean relation from Croton et al. (2006) as in Baldry et al. (2006) (although as they note, the relation has considerable scatter). Similarly, we show the poststarburst (generally merger remnant) fraction from Hogg et al. (2006) and Goto (2005) as a function of surface density on large scales.

Our predictions and the observations emphasize that galaxy mergers occur on all scales (in halos of all masses) and in all environments. In a global sense, there is no preferred merger scale. That is not to say that mergers of galaxies of a particular mass do not have a preferred scale (indeed, in our modeling, this is explicitly the small group scale), but rather because this scale is a function of galaxy mass, mergers of *some* mass occur in all halo masses and environments. It is clear that it is a mistake to think that mergers would not occur in field (or even void) environments, a fact that is very important to the formation of spheroids and quasars in these locations.

2.3. *How Mergers Are Influenced by Environment*

Figure 6 demonstrates that, all else being equal, mergers do not depend on the large-scale environment. This is conventional wisdom, of course, because mergers are an essentially *local*
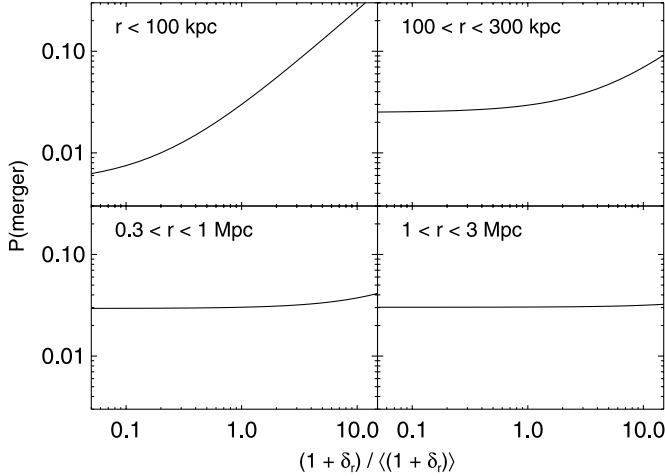
Fig. 7.—Dependence of the merger rate/probability on environmental density decrement/enhancement within a given radius r; i.e., galaxy overdensity $(1 + \delta_r)/\langle(1 + \delta_r)\rangle$ at a fixed galaxy and host halo mass (absolute units are arbitrary here, and depend on these quantities). On scales less than the typical virial radii of interest, the merger rate increases with overdensity (linearly at $\delta_r \gg 1$), but it is independent (for a fixed halo mass) of large-scale environment.

process. However, there is one sense in which the merger rate should depend on environment. If the local density of galaxies (supply of systems for major mergers) is enhanced by some factor $1 + \delta$, then the probability (or rate) of major mergers should be enhanced by the same factor.

In detail, our adopted model for the merger/capture cross section of galaxies (§ 2.1) allows us to calculate the differential probability that some halo/subhalo or galaxy population at a given distance r will merge with the central galaxy in a time $< t_H$. Given the observed galaxy-galaxy correlation function as a function of stellar mass (Li et al. 2006), we can trivially calculate the mean number density of galaxies (possible fuel for major mergers) in a shell dr at r, and combining this with the merger rate/cross section calculation determines the differential contribution to the total merger rate of galaxies of that mass, from pairs at the separation dr. This can be thought of as either a capture process from halo/subhalo orbits or a global inflow rate from dynamical friction and gravitational motions; the results are the same, modulo the absolute merger rate normalization (Binney & Tremaine 1987; Masjedi et al. 2006). Next, assume that the density of these companions is multiplied, at this radius, by a factor $1 + \delta_r$ (relative to the mean $\langle(1 + \delta_r)\rangle$ expected at that r for the given central halo mass). Integrating over all radii, we obtain the total merger rate/probability, with the appropriate enhancement.

Figure 7 illustrates this, calculated in several radial shells using our gravitational capture cross sections to estimate the enhancement (the other cross sections yield similar results). The absolute value of the probability shown will be a function of galaxy mass, halo mass, and redshift, but the qualitative behavior is similar. Unsurprisingly, density enhancements on small scales ($r \lesssim 100$ kpc, where most systems will merge) linearly increase the merger rate accordingly. Note that density decrements decrease the merger rate only to a point—this is because even for a galaxy with no companions within a 100 kpc radius, there is of course some nonzero probability that companions will be accreted or captured from initially larger radii and merge in $t \ll t_H$.

At larger radii, the enhancement is less pronounced. A galaxy in the center of a halo of a given mass in a ~ 3 Mpc overdensity is not substantially more likely to experience a major merger, because there is little contribution to its merger rate from those large radii (at least on short timescales; of course, over $t \sim t_H$ subhalos

may be accreted from these radii, but by then the density structure will change and the merger rate will reflect that). Naturally, an overdensity at the ~ 3 Mpc scale implies an enhanced density within that scale. However, we are considering this for galaxies and halos of a specific mass, for which the virial radii are generally much smaller than these scales, so the increased density in this annulus does not necessarily imply an enhanced galaxy density within the halos themselves (for that M_{halo}), although it may affect the overall abundance of the halos. As a general rule, merger rates will scale with environmental density on scales less than the virial radii of the masses of interest and be independent of density on larger scales.

If the merger rate increases in regions with small-scale overdensities, then mergers themselves should be biased to such regions. To the extent that the small-scale galaxy overdensity around a merger traces this overdensity (which we caution is not *necessarily* true, as one of the initial galaxies in this overdensity is, by definition, consumed in the merger), this implies that mergers and merger remnants should preferentially exhibit small-scale density excesses. The magnitude of this excess is straightforward to determine: for a given galaxy/halo mass, the distribution of environments [densities $(1 + \delta_r)$ on a given scale r] is known. Then, for each scale r, the calculation in Figure 7 gives the relative probability of a merger as a function of overdensity. Convolving the probability of any object being in given overdensity with the probability of a merger in that overdensity gives the mean overdensity of mergers at that scale, i.e.,

$$\frac{\langle x_{merger}\rangle}{\langle x_{all}\rangle} = \frac{\int x P_{merger}(x) P(x|M_{halo}) dx}{\int x P(x|M_{halo}) dx}, \qquad (2)$$

where $x \equiv (1 + \delta_r)$.

It is straightforward in extended Press-Schechter theory to calculate of the probability of forming a halo of a given mass in a given overdensity on a particular scale (Mo & White 1996). However, since we are calculating a galaxy overdensity in radii about the merger candidate, Poisson noise is dominant on small scales where the average number of companions is $\lesssim 1$—nevertheless, it is again straightforward to calculate the probability of a given overdensity. In any case, we account for both effects and show the results in Figure 8. Specifically, we show the average number of companions within a radius of a given r about a merger, for all field galaxies. We then multiply the field curve by the calculated overdensity of mergers as a function of r. The exercise can then be trivially repeated for the correlation function $\xi(r)$. We compare with observed poststarburst populations (E+A/K+A) galaxies and find that they display a similar excess on small scales. As before, the difference on large scales is negligible—unsurprisingly, the density excess becomes important at $r \lesssim r_{vir}$ for the typical galaxies of interest.

Finally, we stress that the excess of companions on small scales does *not*, in this model, stem from those galaxies themselves having any interaction with the central merger (remnant) but reflects a genuine small-scale overdensity (as in small groups), in which mergers will be more likely.

2.4. Integrated Merger Populations over Time

At a given redshift, we use our model to predict the mass function of mergers. For clarity, we take the mass of a merger to be the total stellar mass of the remnant galaxy (roughly the total baryonic mass of the merger progenitors). This avoids ambiguity in merger mass ratios, tends to be observationally representative (since mergers are generally labeled by total luminosity/stellar mass), and has been shown in simulations to be a better proxy
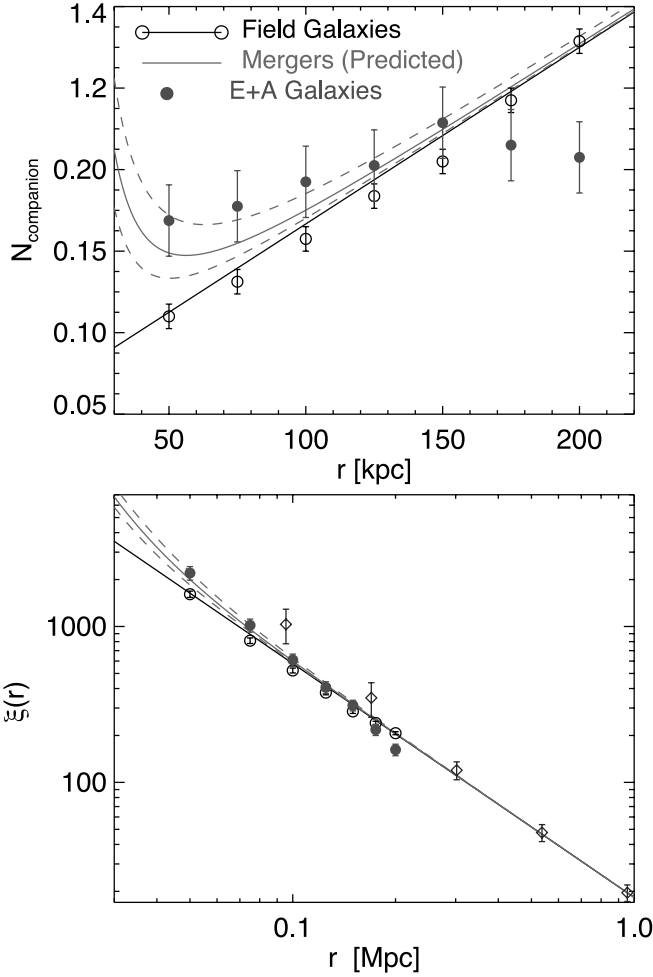
FIG. 8.—Excess galaxy overdensity on small scales predicted for mergers from our model. Because mergers are more likely when there is a galaxy overdensity on small scales (Fig. 7), mergers will, on average, occur in regions with slightly enhanced small-scale densities. We show the real-space correlation function (*bottom*; technically the merger-galaxy cross correlation function) and corresponding number of companions within a given radius (*top*) of all field galaxies (Goto 2005), and then this is multiplied by the predicted excess on small scales for mergers (essentially integrating over the probability bias to large overdensity on small scales in Fig. 7). Dashed gray lines indicate the errors in our estimate from the combination of uncertainties in the field galaxy correlation function, the range of galaxy masses considered (which slightly shifts the physical scale on which the effect is important), and the inclusion/exclusion of Poisson noise in the distribution of overdensities for a given halo mass. The observed number of companions and clustering of poststarburst (likely merger remnant) galaxies is shown for comparison, from Goto (2005, *gray circles*) and Hogg et al. (2006, *black diamonds*). [*See the electronic edition of the Supplement for a color version of this figure.*]

for the merger behavior than the initial mass of either progenitor (as long as it is still a major merger; Hopkins et al. 2006a).

Figure 9 shows the mass functions of ongoing mergers at each of several redshifts. We first consider the mass function of "all" objects that will merge efficiently—i.e., the mass function of merging pairs. This requires no knowledge of the actual timescale of the merger or, e.g., lifetime of tidal disturbances. The results agree well with the mass functions and merger fractions estimated at all $z \lesssim 1.5$, suggesting that our model does indeed reasonably describe the true nature of galaxy mergers. For comparison, we show the results obtained using a different halo occupation model to associate galaxies and halos or using a different set of simulations/models to estimate the subhalo mass functions. As noted in § 2.1, these choices make very little difference (considerably smaller than, e.g., the systematics in the observations).

It is not always clear, however, that observations capture all merging pairs (or that our definition of "all" is appropriate as, for some mergers, $t_{\mathrm{merger}} \to t_{\mathrm{H}}$). Often, systems are identified as mergers on the basis of tidal disturbances and other clear morphological signatures. We therefore calculate the mass function of systems observed in this manner. This requires that we adopt one of the models in § 2.1 for the merger timescale, which tells us how long it will characteristically take for a given merger to reach the interaction cross section where tidal disturbances will be excited. Then, using numerical simulations to estimate the typical duration of those features (in which they will be identified by typical morphological classification schemes; see J. M. Lotz et al. 2008, in preparation), we obtain the observed "disturbed morphology" mass functions. We perform this calculation using each of the methods for calculating the merger timescale described in § 2.1. Note that the number of systems according to this convention can exceed that in our "all pairs" definition if the timescale on which disturbances are visible is longer than the "infall" timescale or timescale on which the subhalo survives (the case for very efficient infall/capture).

At high masses, the difference between samples of merging pairs and those of disturbed systems is small, as is the difference between our choice of methodology in calculating the merger abundances and/or timescales. This is because high-mass systems merge more quickly, excite morphological disturbances more easily on first passage, and are brighter (making faint morphological features easier to identify). At very low masses $M_{\mathrm{gal}} \lesssim 10^{10}\ M_{\odot}$, our predictions do diverge—this is because the overall infall or merger timescale can become substantially longer than the timescale over which morphological disturbances are excited (in these cases, this occurs closer to the final coalescence). Although this conclusion merits more detailed numerical investigation in future work, it has little effect on any of our predictions—for example, the total merger fraction (especially at high redshift) is restricted to larger mass $M_{\mathrm{gal}} \gtrsim M_*$ systems, where the predictions agree well, and the overall merger mass density is nearly identical regardless of the methodology. Furthermore, quasar and galaxy formation processes are probably influenced (or even dominated) by other mechanisms (such as secular disk instabilities and quenching via infall as a satellite galaxy) at these low masses, which we do not attempt to model.

We next integrate the mass functions in Figure 9 above a given mass limit to predict the merger fraction as a function of redshift, shown in Figure 10. The fraction is determined relative to the mass functions in Fontana et al. (2006), who provide a continuous fit over the range of interest. But we note that since this is an integrated quantity, the difference adopting other mass function estimates (e.g., Borch et al. 2006) is small (at least at $z \lesssim 1.5$). Comparing this to a range of observations, the agreement is good, especially for the deeper mass limit. For high-mass mergers ($M_{\mathrm{gal}} \gtrsim 10^{11}\ M_{\odot}$) there is greater scatter in the observations, which most likely is due to cosmic variance (especially at $z \lesssim 0.2$). In both cases, however, the merger fraction is not an especially steep function of redshift. In fact, for $z = 0.3-1.5$, the fraction increases by only a factor of $\sim 3-4$, consistent with most observations finding a relatively flat merger fraction in this range (e.g., Lin et al. 2004; Lotz et al. 2008) and recent cosmological simulations (Maller et al. 2006). Further, although halos may be merging more frequently at high redshift, they are also merging more rapidly, meaning that the fraction merging at any instant can be relatively flat.

Finally, given our model for the halos hosting mergers, it is straightforward to calculate the predicted clustering properties of those mergers. Specifically, we have already predicted a number density of mergers as a function of halo mass, galaxy mass, and
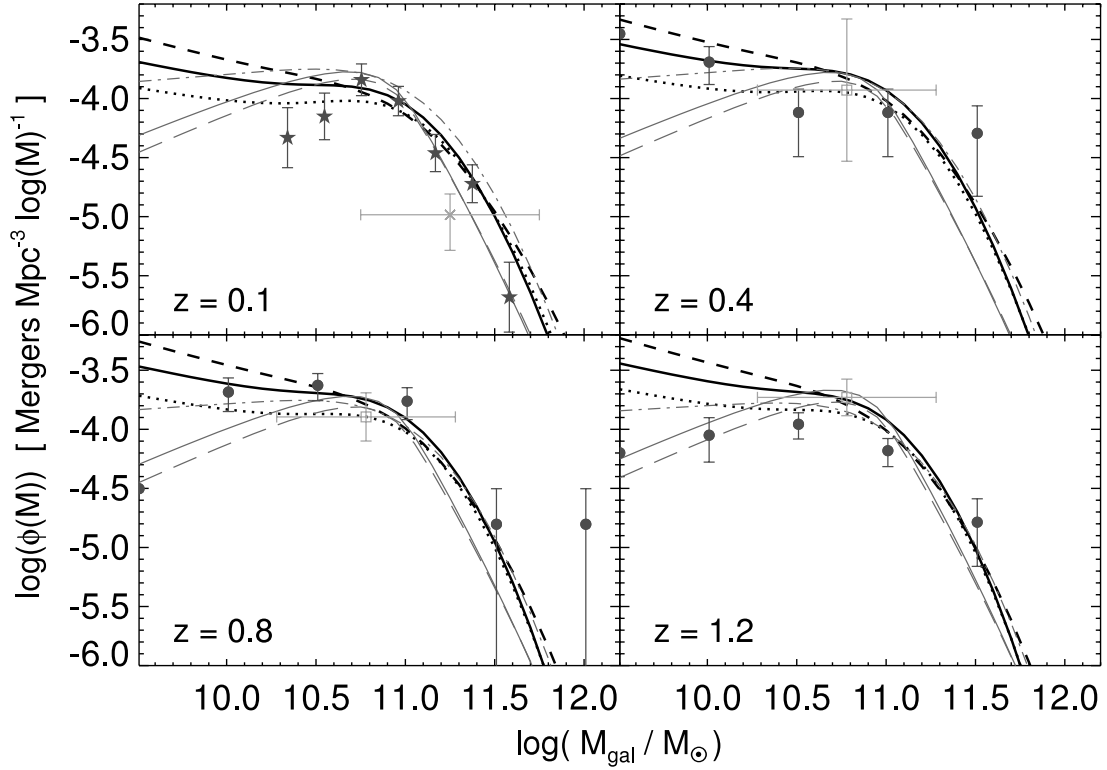
Fig. 9.—Mass functions (in terms of the remnant stellar mass) of ongoing mergers at each of several redshifts (labeled). Observed mass functions (*solid gray points*) are shown from Xu et al. (2004, *stars*) and Bundy et al. (2005, *circles*; for a detailed analysis of the mass functions, see Hopkins et al. 2007a). Error bars do *not* include cosmic variance. Observed merger fractions (*open gray points*), converted to a mass function estimate over the mass range sampled (*horizontal errors*) are shown from Bell et al. (2006, *cross*) and Lotz et al. (2008, *squares*), with errors including cosmic variance. We compare the prediction of our default model (*thick solid black line*) for the abundance of mergers and merging pairs. Dotted line employs a different halo occupation model, and dashed line adopts a different fit to the subhalo mass functions (see Fig. 6 and § 2.1.2). We also show the predictions for morphologically identified mergers (*thin gray lines*), which requires estimating the merger timescale/capture efficiency and duration of morphological disturbances (see § 2.1). We estimate these using a group capture/collisional model (*solid line*), angular momentum capture cross sections (*long-dashed line*), and simple dynamical friction considerations (*dotted line*), calibrating the duration of disturbances from numerical simulations (J. M. Lotz et al. 2008, in preparation). At masses $\gtrsim 10^{10} M_\odot$, there is little difference due to methodology. At very low masses, simulations suggest that the merger timescale (i.e., orbital or crossing time after first passage) is considerably longer than the time period over which strong disturbances are excited; however, this is below the mass scales of interest for most of our predictions. [*See the electronic edition of the Supplement for a color version of this figure.*]
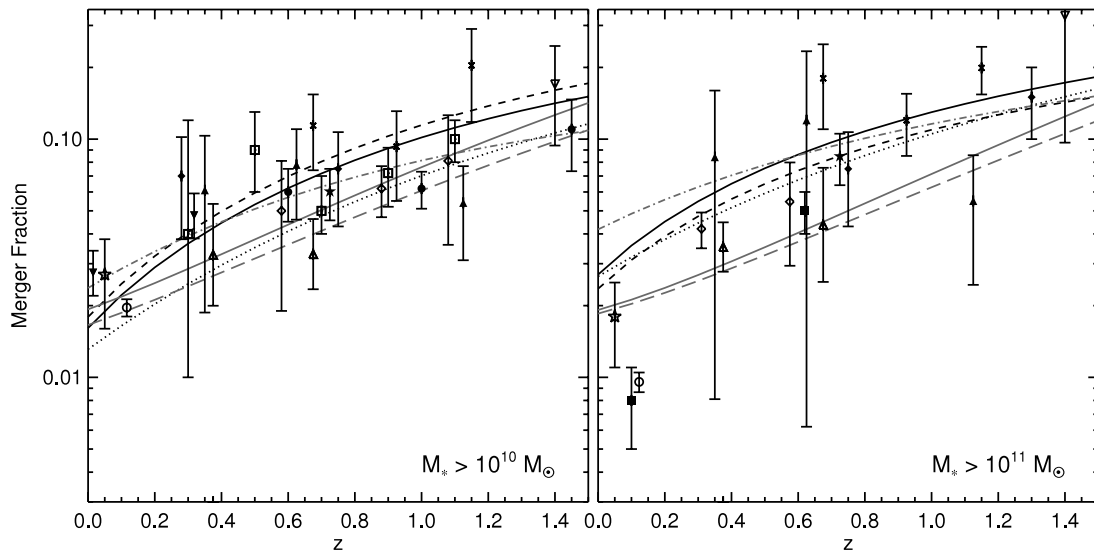

Fig. 10.—Predicted merger fraction as a function of redshift (*lines*, same style as Fig. 9), above two approximate mass limits. Observations (*points*) are shown from Patton et al. (2002, *filled inverted triangles*), Conselice et al. (2003, *filled circles*), Bundy et al. (2004, *filled triangles*) Lin et al. (2004, *open diamonds*), Xu et al. (2004, *open stars*), De Propris et al. (2005, *open circles*), Cassata et al. (2005, *filled diamonds*), Wolf et al. (2005, *filled stars*), Bundy et al. (2005, *open triangles*), Lotz et al. (2006, *open inverted triangles*), Lotz et al. (2008, *open squares*), Bell et al. (2006, *filled squares*), and Bridge et al. (2007, *crosses*). Note that the mass limit is only approximate in several of these cases, as they are selected by optical luminosity. The predicted merger fractions agree well, especially for the deeper case, which resolves M_* galaxies. [*See the electronic edition of the Supplement for a color version of this figure.*]
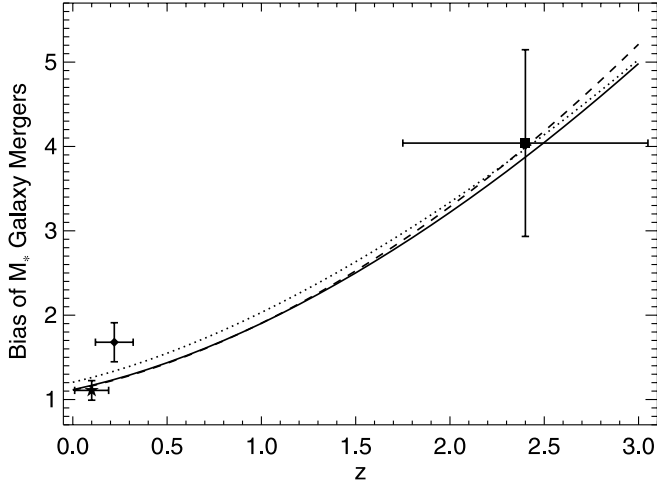
FIG. 11.—Comparing our predicted clustering of $\sim M_*$ major mergers (*lines*; style as in Fig. 9) as a function of redshift to that various populations usually associated with galaxy mergers (*points*): poststarburst (E+A/K+A) galaxies (Blake et al. 2004, *star*), close galaxy pairs (Infante et al. 2002, *diamond*), and submillimeter galaxies (Blain et al. 2004, *square*).

redshift; i.e., some $n_{\mathrm{merger}}(M_{\mathrm{gal}}|M_{\mathrm{halo}}, z)$. Knowing the clustering amplitude or bias of each host halo $b(M_{\mathrm{halo}}|z)$, it is straightforward to predict the clustering of the merging galaxies, in the same manner by which halo occupation models construct the clustering of a given population:

$$b(M_{\mathrm{gal}}) = \frac{\int b(M_{\mathrm{halo}}) n_{\mathrm{merger}}(M_{\mathrm{gal}}|M_{\mathrm{halo}}) dM_{\mathrm{halo}}}{\int n_{\mathrm{merger}}(M_{\mathrm{gal}}|M_{\mathrm{halo}}) dM_{\mathrm{halo}}}. \qquad (3)$$

We calculate $b(M_{\mathrm{halo}})$ following Mo & White (1996) as updated by Sheth et al. (2001) to agree with the results of numerical simulations.

Figure 11 shows this as a function of redshift. Since observations generally sample near $M_{\mathrm{gal}} \sim M_*$, we plot this for $M_{\mathrm{gal}} = M_*(z=0) \approx 10^{11}\ M_\odot$. We compare with available clustering measurements for likely major-merger populations. At low redshifts, Blake et al. (2004) have measured the clustering of a large, uniformly selected sample of poststarburst (E+A/K+A) galaxies in the 2dF. Infante et al. (2002) have also measured the large-scale clustering of close galaxy pairs selected from the SDSS at low redshift. At high redshift, no such samples exist, but Blain et al. (2004) have estimated the clustering of a moderately large sample of spectroscopically identified submillimeter galaxies at $z \sim$ 2–3, which as discussed in § 1 are believed to originate in major mergers. Our prediction is consistent with these constraints— however, given the very limited nature of the data and the lack of a uniform selection criteria for ongoing or recent mergers at different redshifts, we cannot draw any strong conclusions.

One caution should be added: recent higher resolution simulations suggest that the approximation here (and in many—but not all—halo occupation models), that bias is a function only of halo mass at a given redshift, may not be accurate (e.g., Gao & White 2007; Harker et al. 2006; Wechsler et al. 2006). In particular, because mergers have particularly recent halo assembly times for their postmerger masses, they may represent especially biased regions of the density distribution. Unfortunately, it is not clear how to treat this in detail, as there remains considerable disagreement in the literature as to whether or not a significant "merger bias" exists (see, e.g., Kauffmann & Haehnelt 2002; Percival et al. 2003; Furlanetto & Kamionkowski 2006; Lidz et al. 2008. Furthermore, the distinction between galaxy-galaxy and halo-halo merg-

ers (with the considerably longer timescale for most galaxy mergers) means that it is not even clear whether or not, after the galaxy merger, there would be a significant age bias.

In any case, most studies suggest the effect is quite small: using the fitting formulae from Wechsler et al. (2002, 2006) we find that even in extreme cases (e.g., a $M_{\mathrm{halo}} \gg M_*$ halo merging at $z = 0$ as opposed to an average assembly redshift $z_f \approx 6$) the result is that the standard EPS formalism underestimates the bias by $\approx 30\%$. For the estimated characteristic quasar host halo masses and redshifts of interest here, the maximal effect is $\lesssim 10\%$ at all $z = 0$–3, much smaller than other systematic effects we have considered (and generally within the range of our plotted variant calculations in Fig. 11). This is consistent with Gao & White (2007) and Croton et al. (2007), who find that assembly bias is only important (beyond the 10% level) for the most extreme halos or galaxies in their simulations, where for example the clustering of small halos that are destined to be accreted as substructure in clusters ($\gtrsim 10^{15}\ h^{-1}\ M_\odot$) will be very different from the clustering of similar-mass halos in field or void environments. Indeed, our own calculation in Figure 8 suggests that merger bias applies only on small scales and that mergers show no preference for excess densities on the large scales for which the linear bias description is meaningful. The effect may grow with redshift, however, so care should be taken in extrapolating the predictions in Figure 11 to higher redshifts. For further discussion of the effects on the data and predictions shown here, we refer to Hopkins et al. (2007d).

For the sake of future comparison, we show in Figure 12 our predictions for the merger fractions and clustering of mergers (Figs. 10 and 11, respectively) at all redshifts $z = 0$–6. We note the caveat that our merger fraction is defined relative to the mass functions in Fontana et al. (2006), which become uncertain at high redshifts, although this uncertainty is comparable to the differences between the methods of calculating the merger timescale (as discussed in § 2.1.2). It is also less clear what the observable consequences of mergers at the highest redshifts may be—if merger rates are sufficiently high, there may be a large number of multiple mergers (as in Li et al. 2007), or systems may effectively be so gas-rich that merging preserves disks and operates as a means of "clumpy accretion" (e.g., Robertson et al. 2006a).

Although the estimates differ at the highest redshifts, we stress that their integrated consequences at low redshifts $z \lesssim 3$ are similar, as this is where most merging activity and spheroid/BH mass buildup occurs. We also note that high-redshift mergers are likely to be the most massive $M_{\mathrm{gal}} \gg M_*$ systems, so we show our predictions for the clustering of mergers assuming different mass limits (as opposed to strictly at $M_{\mathrm{gal}} = M_*$). We also plot the mass flux in mergers, i.e., the integrated rate at which galaxy baryonic/stellar mass is merged, $\int M_{\mathrm{gal}} \dot{n}(M_{\mathrm{gal}}) d \log M_{\mathrm{gal}}$. This compares favorably with the observationally inferred rates at which mass is moved off the blue cloud, through the "green valley," and onto the red sequence (from the evolution in galaxy mass functions and color-magnitude relations; see Martin et al. 2007), as expected in a model where mergers drive such a transition (for details, see Paper II). Future observations of these quantities at high redshift will improve the constraints on our halo occupation and merger time-scale estimates, allowing for more accurate calculations of, e.g., quasar triggering and spheroid formation rates at these epochs.

3. QUASARS

3.1. *Consequences of Merger-driven Fueling: What Determines Where and When Quasars Live*

Having developed in § 2 a physically motivated model of merger rates as a function of galaxy and halo mass, environment,
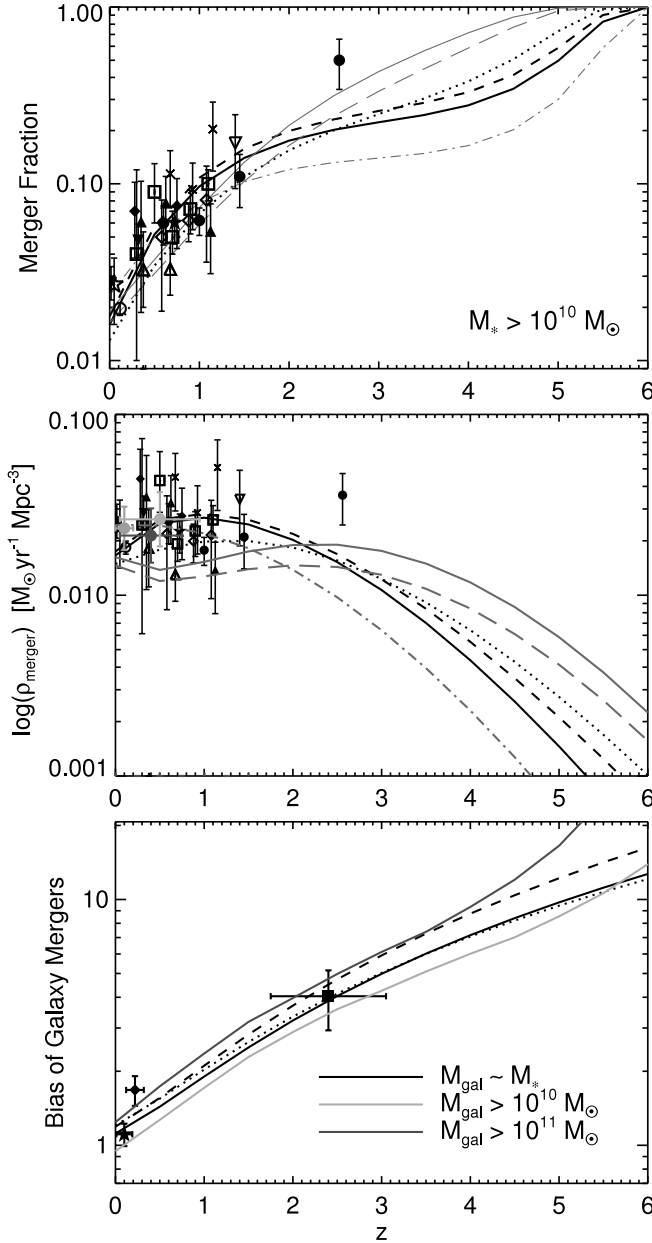
FIG. 12.—*Top*: Same as Fig. 10, but extending our predicted merger fractions to high redshift. *Middle*: Mass flux through mergers (i.e., total rate of stellar mass merging). Black points are observed merger fractions converted to an estimated mass flux rate following Hopkins et al. (2007a). Gray circles show the observationally inferred mass flux through the "green valley" (i.e., from blue cloud to red sequence), rate of growth of the red sequence, and rate of mass loss off the blue cloud (respectively), from $z \sim 0-1$ (Martin et al. 2007; see Paper II for a more detailed comparison). *Bottom*: Same as Fig. 11, but extended to higher redshift. Gray lines show the clustering of mergers above the given mass thresholds. [*See the electronic edition of the Supplement for a color version of this figure.*]

and redshift (and having tested that this model is consistent with the existing body of merger observations), we can now extend our application. As discussed in § 1, the argument for an association between mergers and quasars has a long history. We therefore make the simple *Ansatz*: *Every major merger of star-forming/ gas-rich galaxies triggers a quasar.*

From this statement, we can make a number of robust predictions. In § 2 we derived the characteristic host halo mass for mergers of $\sim M_*$ galaxies. To the extent that these are gas-rich systems, this should therefore also represent the characteristic

host halo mass of quasars and (since the mass density of the universe is dominated by systems near $\sim M_*$) dominate the buildup of black hole mass.

From the Soltan (1982) argument, the black hole mass density of the universe must be dominated by growth in typical, bright quasar phases with canonical radiative efficiency $\epsilon_r \sim 0.1$. Let us construct the simplest possible model: mergers (of M_* galaxies) characteristically occur at a host halo mass $\sim M_{\mathrm{merger}}$. From the halo mass function, it is straightforward to calculate the rate at which halo mass crosses this mass threshold:

$$\dot{\rho}_{\mathrm{halo}} = \bar{\rho} \frac{dF(> M_{\mathrm{halo}})}{dt}, \qquad (4)$$

where $F(> M_{\mathrm{halo}}, z)$ is the fraction of mass in halos of mass greater than M_{halo}, determined from the Press-Schechter formalism revised following Sheth et al. (2001). Assume that every such halo undergoes a merger approximately on crossing this mass threshold, which transforms its galaxy from disk to spheroid. The hosted BH mass therefore grows from some arbitrarily small amount to the expected mass given the BH-host mass relations, which we can write as $M_{\mathrm{BH}} = \nu(z) M_{\mathrm{halo}}$ [we distinguish this from $M_{\mathrm{BH}} = \mu(z) M_{\mathrm{gal}}$]. The ratio $\nu(z)$ is determined to $z \sim 3$ from the clustering of active BHs of a given mass at each redshift (see, e.g., da Angela et al. 2007; Fine et al. 2006; Hopkins et al. 2007d, 2007c) and indirectly from determinations of the BH host galaxy masses (Peng et al. 2006). The total rate at which BH mass is built up is then

$$\dot{\rho}_{\mathrm{BH}} = \nu(z) \dot{\rho}_{\mathrm{halo}} = \nu(z) \bar{\rho} \frac{dF(> M_{\mathrm{halo}})}{dt}, \qquad (5)$$

and the bolometric luminosity density is $j_{\mathrm{bol}} = \epsilon_r \dot{\rho}_{\mathrm{BH}} c^2$. Figure 13 compares this simple estimate with the observed bolometric quasar luminosity density as a function of redshift.

The agreement is striking, which suggests that this toy model, such that the bulk of the assembly of BH mass occurs near the transition halo mass, is reasonable. This also naturally explains the rise and fall of the quasar luminosity density with time. However, this is ultimately just a simple approximation—we can consider this in greater detail adopting our previous estimate of the merger rate as a function of stellar mass and redshift, $\dot{n}(M_{\mathrm{gal}}|z)$, from § 2. Each major merger transforms disks to spheroids, building a BH of average mass $M_{\mathrm{BH}} = \mu(z) M_{\mathrm{gal}}$. We should properly only consider mergers of gas-rich or star-forming systems, as dry mergers will, by definition, not be able to trigger quasar activity and form new BH mass. Therefore, we empirically adopt the fraction of red and blue galaxies at each M_{gal}, M_{halo} (as in § 2) to restrict only to mergers of blue galaxies. Again, $\mu(z)$ has been directly determined from observations (Peng et al. 2006) and estimated from theoretical arguments (Hopkins et al. 2007c). For convenience, we adopt the numerical best-fit estimate of $\mu(z)$ from Hopkins et al. (2007c). A good approximation to this numerical function is

$$\mu(z) \approx 0.0012 \left[\frac{1 + z^{5/2}}{1 + (z/1.775)^{5/2}} \right], \qquad (6)$$

which matches the asymptotic observed values at low and high redshift (Häring & Rix 2004; Walter et al. 2004) and captures the observed weak evolution to $z \sim 1$ and rapid evolution between $z = 1-3$ (Shields et al. 2003; Peng et al. 2006; Salviander et al. 2006). Given the merger rate $\dot{n}(M_{\mathrm{gal}}|z)$, we can then convert
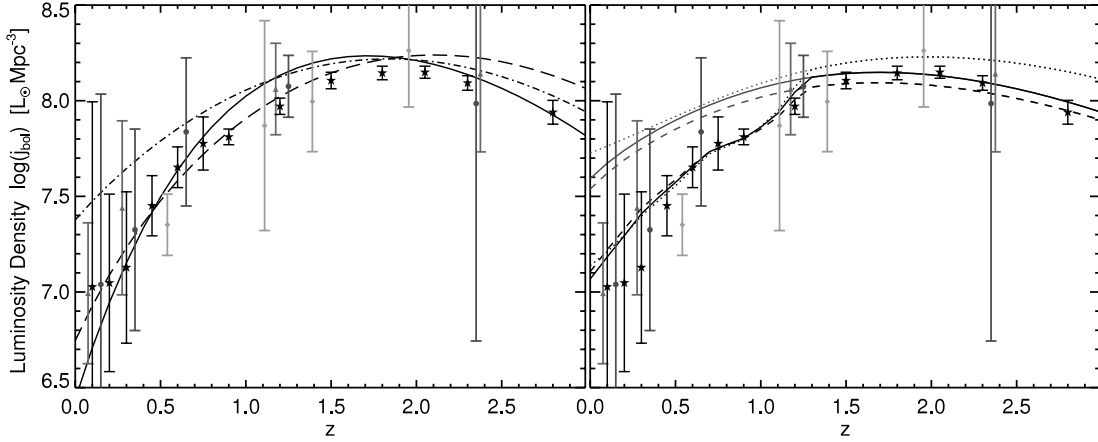
FIG. 13.—Predicted quasar luminosity density, if quasars are triggered in mergers, as a function of redshift. *Left*: Prediction from a simplified toy model in which all halos hosting $\sim L_*$ galaxies undergo major mergers near their characteristic small group mass scale and build a BH that obeys the appropriate $M_{\rm BH}$-$M_{\rm halo}$ relation for that redshift (estimated $M_{\rm BH}$-$M_{\rm halo}$ as a function of redshift from Hopkins et al. 2007d, Fine et al. 2006, and Hopkins et al. 2007c corresponding to solid, long-dashed, and dot-dashed lines, respectively). Points show observational estimates from the measured QLFs of Ueda et al. (2003, *gray circles*), Hasinger et al. (2005, *gray triangles*), Richards et al. (2005, *gray diamonds*), and the large compilation of multiwavelength QLF data in Hopkins et al. (2007e, *black stars*). The observations from specific bands are converted to a bolometric luminosity density using the bolometric corrections calibrated in Hopkins et al. (2007e). *Right*: Same, but the predicted luminosity density is calculated properly accounting for all galaxy and halo masses from the merger rate functions determined in § 2 and adopting the observed ratio of BH to host galaxy spheroid mass as a function of redshift (e.g., Peng et al. 2006). Line styles correspond to different means of estimating the exact merger rates, as in Fig. 9. Red lines assume all mergers will trigger quasars, black (*lower*) lines assume only gas-rich ("wet") mergers can trigger bright quasar activity (adopting the observed fraction of gas-rich/star-forming/blue galaxies as a function of $M_{\rm gal}$ and $M_{\rm halo}$ as in Fig. 5). A merger-driven model naturally predicts both the rise and fall of the global quasar luminosity density to high precision. [*See the electronic edition of the Supplement for a color version of this figure.*]

this to a cosmic rate of formation or buildup of BHs in merger-driven quasars:

$$\dot{n}(M_{\rm BH}|z) = \int P(M_{\rm BH}|M_{\rm gal})\dot{n}(M_{\rm gal}|z)d\log M_{\rm gal}. \qquad (7)$$

The intrinsic dispersion about the mean BH-host mass relation appears, at all redshifts, to be roughly lognormal with width ≈ 0.27 dex, so we model $P(M_{\rm BH}|M_{\rm gal})$ as such. Once the total rate of formation of BH mass is calculated, the same conversion above yields the quasar luminosity density.

Figure 13 shows the results of this more detailed calculation. They are similar to the results from our extremely simplified model—which reflects the fact that most of the mass/luminosity density is contained near M_* or L_*. Note that considering all mergers (i.e., including dry mergers) overpredicts the quasar luminosity density at low redshifts. This demonstrates that the decrease in the quasar luminosity density at low redshifts is, in part, driven by the fact that an increasing fraction of massive systems have already been transformed to "red and dead" systems at late times and are no longer available to fuel quasars, even if they undergo subsequent dry mergers. By $z \sim 0$, for example, a large fraction ($\sim 50\%$) of the mass density in $>M_*$ systems has already been gas-exhausted (discussed in detail in Paper II), and therefore such mergers are no longer a viable fuel supply for quasar activity. As discussed in § 2.2, the predicted gas-rich merger mass density (and corresponding quasar luminosity density) at $z \lesssim 0.5$ will be slightly lower if these gas-exhausted systems are preferentially surrounded by gas-exhausted satellites (compared to gas-rich central galaxies of the same mass in similar halos), but it is clear in Figure 13 that this is completely consistent with the observations (especially if secular processes contribute significantly to the quasar luminosity density at low redshifts and luminosities, as we expect from our comparisons in § 3.3).

Having calculated the rate of BH formation as a function of the remnant BH mass, $\dot{n}(M_{\rm BH}|z)$, it is trivial to integrate this forward and predict the BH mass function (BHMF) at any time. Fig-

ure 14 shows the result of this calculation at $z = 0$, compared to the observationally estimated BHMF. The two agree well at all masses, even at very large $M_{\rm BH} \sim 10^{10}\, M_\odot$. We also show the BHMF at several other redshifts. Interestingly, there is a downsizing behavior, where a large fraction of the most massive BHs are in place by $z = 2$, while less massive BHs form later (essentially required by the fact that few $\sim 10^9\, M_\odot$ BHs are active at low redshift, while a very high fraction are active at $z \sim 2$; see McLure & Dunlop 2004; Kollmeier et al. 2006; Fine et al. 2006). If we were to ignore dry mergers at low redshifts, this effect would be even
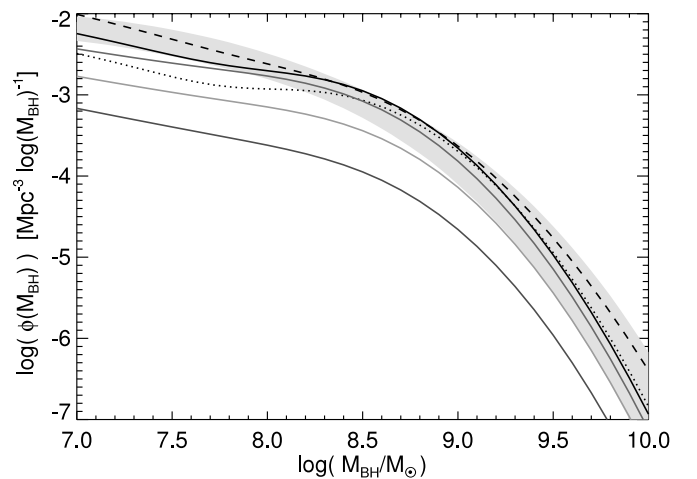

FIG. 14.—Predicted BH mass function (BHMF) from gas-rich merger-driven quasar/BH formation (Fig. 13, *right*). Results are shown at $z = 0$ (*black lines*; line styles correspond to different calculations of the merger rates, as in Fig. 9), and $z = 1, 2, 3$ (*top, middle, and bottom gray lines, respectively*; for clarity, only our fiducial calculation (*solid line*) is shown, but relative evolution with redshift for each calculation is similar). Gray (*shaded*) range shows the $z = 0$ observational estimate of the BHMF in Marconi et al. (2004). Integrating forward the merger mass functions as a function of redshift yields a good match to the local BHMF. The effect of dry mergers is included but is small. [*See the electronic edition of the Supplement for a color version of this figure.*]
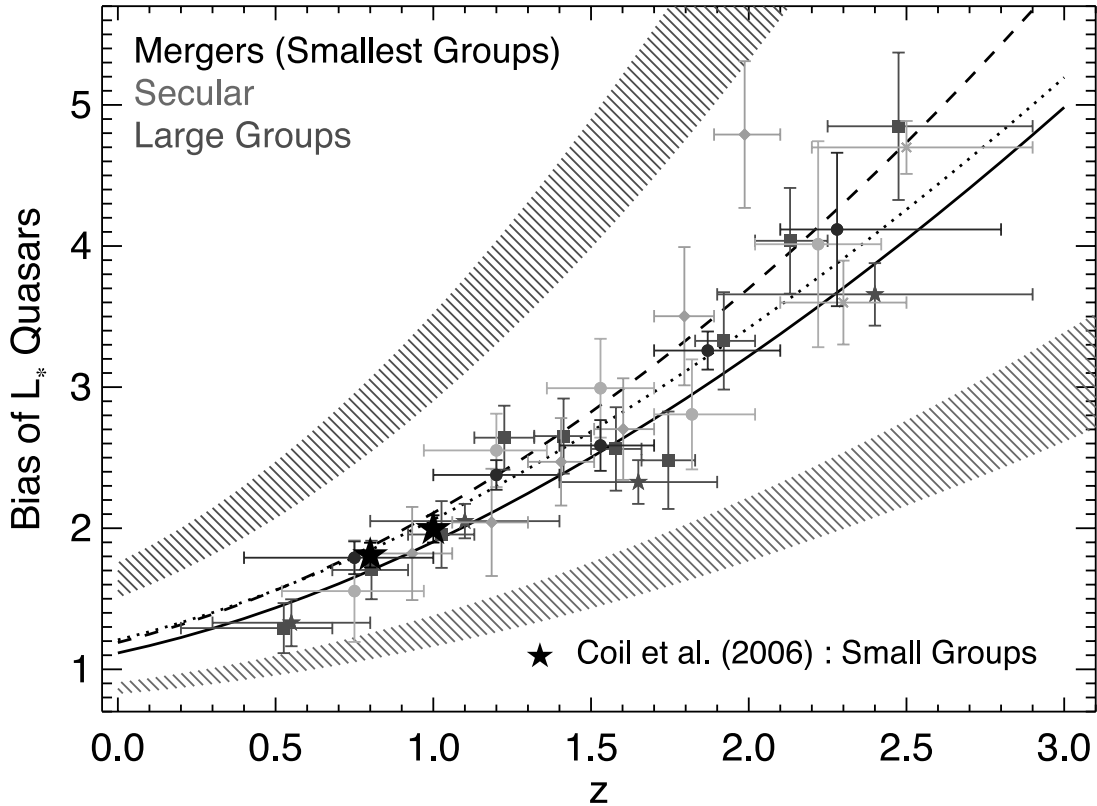
FIG. 15.—Predicted quasar clustering as a function of redshift, assuming merger triggering (*black lines*, as in Fig. 9), corresponding to the small group scale of $\sim M_*$ galaxies. Upper shaded range show the prediction if quasars were associated with large group scales, lower range show the prediction from a secular model in which quasar clustering traces that of star-forming galaxies observed at each redshift (lines show $\pm 1\,\sigma$ range estimated from the compiled observations in Hopkins et al. 2007d from Shepherd et al. 2001, Giavalisco & Dickinson 2001, Norberg et al. 2002, Coil et al. 2004, Zehavi et al. 2005, Adelberger et al. 2005, Allen et al. 2005, Phleps et al. 2006, Meneux et al. 2006, and Lee et al. 2006). Points show quasar clustering measurements from Croom et al. (2005, *gray squares*), Porciani & Norberg (2006, *gray diamonds*), Myers et al. (2007a, 2006, *gray circles*), and da Angela et al. (2007, *gray stars*). Large black stars show the observed clustering of $z \sim 1$ small groups (of $\sim L_*$ galaxies) from Coil et al. (2006) corresponding to the most efficient scales for major $\sim L_*$ galaxy mergers. Quasar clustering measurements are consistent with the small group scale in which mergers proceed efficiently. [*See the electronic edition of the Supplement for a color version of this figure.*]

more pronounced, but at $z \lesssim 1$ their effect is to move some of the BH mass density from lower mass systems into higher mass $\gtrsim 10^9\,M_\odot$ systems (at higher redshifts, the effects are negligible). It is not obvious, however, that this translates to downsizing in galaxy mass assembly, since the ratio of BH to galaxy mass $\mu(z)$ evolves with redshift. We will return to this question in Paper II.

Since we begin our calculation with the halos hosting quasars, we should be able to predict the bias of quasars as a function of redshift. As in Figure 11, we use the known clustering of the halos hosting mergers to calculate the clustering of those mergers as a function of redshift. Assuming each merger produces a quasar of the appropriate mass, this yields the expected clustering of quasars as a function of redshift. Figure 15 compares this prediction to observed quasar clustering as a function of redshift. Technically, we adopt the quasar light curve models from § 3.2 below to determine the clustering specifically of L_* quasars (i.e., determining the relative contribution to L_* from different host masses and their clustering as in Fig. 11), but the result is nearly identical to assuming that L_* quasars trace M_* mergers (Fig. 11). This should be true in any model, as long as the quasar lifetime is a smooth function of luminosity or host mass. We also compare with the directly observed clustering of small groups similar to our definition.

The agreement is quite good at all $z \lesssim 2$. At higher redshifts, the observations show considerably larger scatter, perhaps due to their no longer being complete near the QLF L_*—future observations, sufficiently deep to clearly resolve L_*, are needed to test this in greater detail. We also consider the predicted clustering if

L_* quasars were associated with the large group scale of M_* galaxies (for simplicity we take this to be halo masses $\gtrsim 5$–10 times larger than the small group scale, where our halo occupation model predicts of order $\gtrsim 3$ satellite $\sim M_*$ galaxies), and the expectation from a secular model, in which quasar clustering traces the observed clustering of star-forming galaxies (taken from the observations collected in Hopkins et al. 2007d)—neither agrees with the observations. Note that these estimates may not be applicable to the highest-redshift quasar clustering measurements, where flux limits allow only the most massive $L \gg L_*$ systems to be observed (but see Fig. 12 for how the clustering amplitude varies with merger masses).

We can invert this and compare the empirically determined scales of quasar host systems with the small group scale that should dominate gas-rich $\sim L_*$ galaxy mergers. Figure 16 shows the mean host mass $M_{\rm halo}$, which corresponds to various quasar clustering measurements (i.e., range of $M_{\rm halo}$ for which the expected quasar bias agrees with the observed $\pm 1\,\sigma$ range). We compare this with direct measurements of the halo masses corresponding to small groups of $\sim M_*$ galaxies, determined from both clustering measurements and velocity dispersion measurements of observationally identified groups with dispersions $\sigma \lesssim 200$ km s^{-1}. We can also estimate the appropriate small group scale from the halo occupation formalism.

Specifically, following the formalism of Conroy et al. (2006) if galaxy luminosity/mass is monotonic with subhalo mass (at the time of subhalo accretion), then we can take any galaxy mass
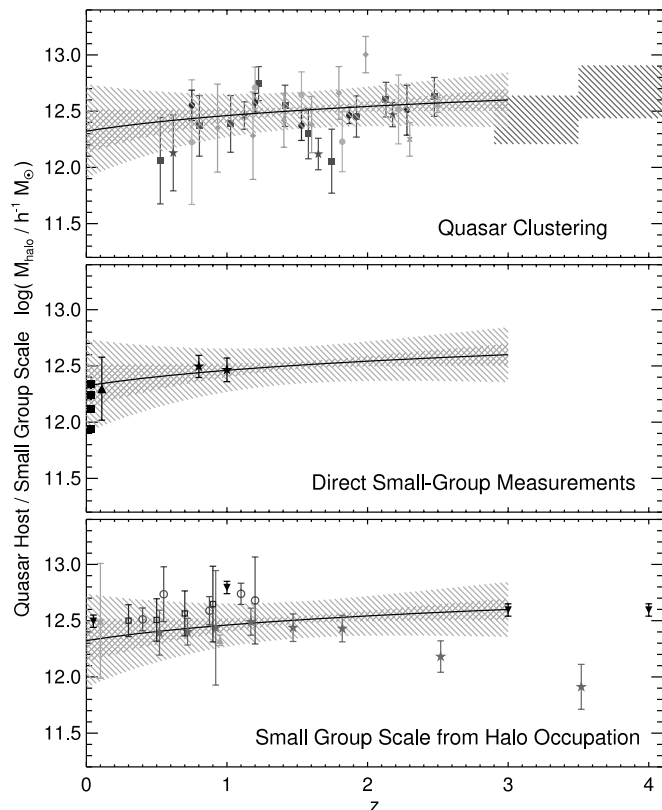
Fig. 16.—*Top*: Characteristic halo mass implied by quasar clustering measurements. Points show the 1 σ allowed range in host halo mass M_{halo} corresponding to the quasar bias measurements in Fig. 15 (in the same style). Shaded regions show the range of halo masses for the corresponding redshift bins in the SDSS (Shen et al. 2007). The solid line shows the best-fit $M_{halo}(z)$ to all observations, with the 1 σ (2 σ) allowed range hatched (cross-hatched). *Middle*: Shaded range again shows the characteristic host halo mass implied by quasar clustering. Points show the halo mass scale implied by direct measurements of observationally identified small groups (velocity dispersions ≲200 km s^{-1}), from Brough et al. (2006) at $z \approx 0$ (*squares*) and from clustering measurements of groups from Eke et al. (2004, *triangles*) and Coil et al. (2006, *stars*). *Bottom*: Same, but showing the small group halo mass estimated indirectly from the empirically determined halo occupation distribution (HOD). Black inverted triangles adopt the best-fit HOD from Conroy et al. (2006) (our default model), other points adopt the methodology of Vale & Ostriker (2006) to construct the HOD from various measured galaxy stellar mass functions in Fontana et al. (2006, *gray stars*), Borch et al. (2006, *gray squares*), Bundy et al. (2005, 2006, *gray circles*), and Blanton (2006, *gray triangles*). The characteristic scale of $\sim L_*$ quasar hosts appears to robustly trace the characteristic small group scale of $\sim L_*$ galaxies—i.e., the mass scale at which galaxy mergers are most efficient. [*See the electronic edition of the Supplement for a color version of this figure.*]

function, monotonically rank it and match it to our halo+subhalo mass functions, and obtain a new halo occupation model that predicts a small group scale—i.e., the range of halo masses at which satellites of mass $\sim L_*$ first appear. As discussed in § 2.1, the choice of mass functions and how the HOD is constructed makes little difference (factor <2) to our predictions, so (unsurprisingly) these all yield a similar estimate of the small group scale to our default model predictions.

At all observed redshifts, the scale of $\sim L_*$ quasars appears to trace the small group scale—i.e., whatever mechanism triggers $\sim L_*$ quasars operates preferentially at the characteristic small group scale for $\sim L_*$ galaxies, where mergers are expected to be most efficient.

In § 2.3, we demonstrated that the increased probability of mergers in regions with excess *small-scale* overdensities means that the typical merger is more likely to exhibit an excess of clustering on small scales, relative to average systems of the same

halo mass. If quasars are triggered in mergers, this should be true as well. We therefore apply the identical methodology from Figure 8 to calculate the excess clustering signal expected in active quasars. Figure 17 shows the results of this exercise. We adopt the large-scale mean clustering expected from Myers et al. (2007b) specifically using the formulae of Smith et al. (2003) to model the expected nonlinear correlation function in the absence of any bias, then we apply the formalism from Figure 8 to estimate the additional bias as a function of scale. Comparing this to observations, the measurements clearly favor an excess bias on small scales ($r \lesssim 100$–$200\ h^{-1}$ kpc; Hennawi et al. 2006), similar to our prediction, over a constant bias at all scales. This appears to be true at all observed redshifts; the excess relative bias we predict at small scales is simply a consequence of how the probability of a merger scales with local density, so it does not vary substantially as a function of redshift.

It should be noted that the excess of quasar clustering on small scales might also reflect an excess of merging binary quasars, i.e., merging systems in which the interaction has triggered quasars in each merging counterpart. For the reasons given in § 1, this situation is expected to be relatively rare (even if all quasars are initially triggered by galaxy mergers), but Myers et al. (2007b) note that only a small fraction of merging pairs need to excite quasar activity in both members in order to explain the observed clustering excess. Figure 17 demonstrates that a similar excess is observed in both the quasar-quasar autocorrelation function (Hennawi et al. 2006; Myers et al. 2007b) and the quasar-galaxy cross-correlation function (Serber et al. 2006), arguing that it primarily reflects a genuine preference for quasar activity in small-scale overdensities. In any case, however, the excess on small scales is a general feature of a merger-driven model for quasar activity. Indeed, the predicted excess is also seen in high-resolution cosmological simulations (Thacker et al. 2006), if quasars are specifically identified with ("attached to") major mergers. Secular (bar or disk instability) fueling mechanisms, on the other hand, should (by definition) show no clustering excess relative to median disk galaxies of the same mass and properties, in contrast to what is observed (although in agreement with what is seen for low-luminosity Seyfert galaxies; see § 3.3).

3.2. Model-dependent Predictions: Additional Consequences of Quasar Light Curves

To proceed further, we must adopt some estimates for quasar light curves and/or lifetimes. Following the methodology developed by Springel & Hernquist (2003) and Springel et al. (2005b), Hopkins et al. (2006a, 2006b) use a large set of several hundred hydrodynamical simulations (see Robertson et al. 2006b) of galaxy mergers, varying the relevant physics, galaxy properties, orbits, and system masses, to quantify the quasar lifetime (and related statistics) as a function of the quasar luminosity. They define the quantity $t_Q(L|M_{BH})$, i.e., the time a quasar of a given BH mass M_{BH} (equivalently, peak quasar luminosity L_{peak}) will be observed at a given luminosity L. They further demonstrate that this quantity is robust across the wide range of varied physics and merger properties; for example, to the extent that the final BH mass is the same, any major merger of sufficient mass ratio (less than $\sim 3:1$) will produce an identical effect. We adopt these estimates in what follows, and note that, while there is still considerable uncertainty in a purely empirical determination of the quasar lifetime, the model light curves are consistent with the present observational constraints from variability studies (Martini 2004 and references therein), clustering (Croom et al. 2005; Adelberger & Steidel 2005; Porciani & Norberg 2006; Myers et al. 2007a; da Angela et al. 2007; Shen et al. 2007), Eddington ratio measurements
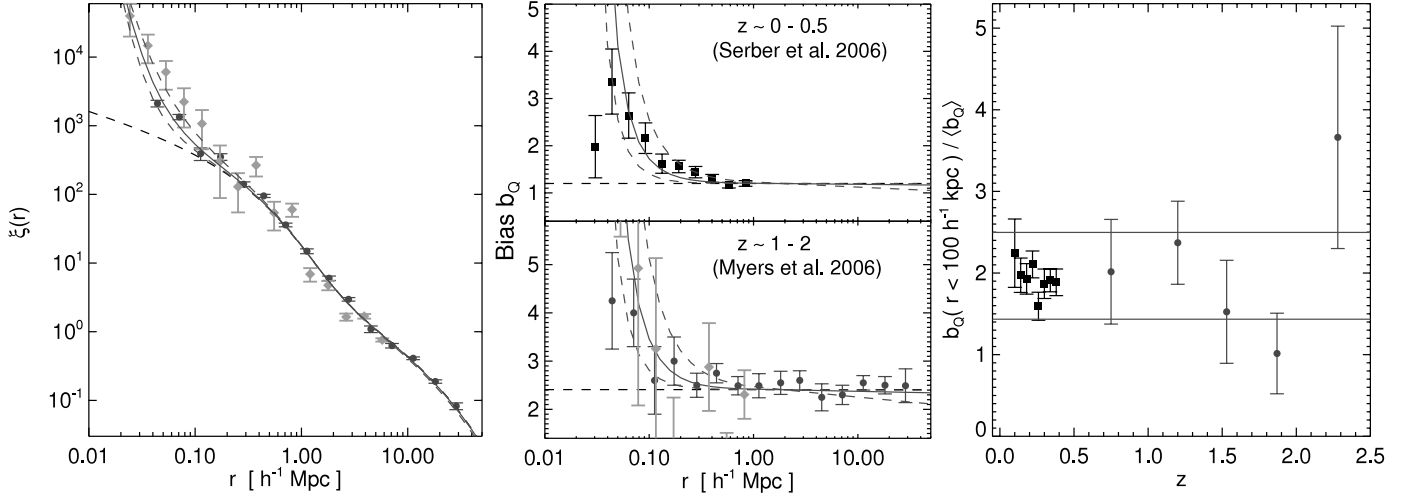
Fig. 17.—Excess small-scale clustering of quasars expected if they are triggered in mergers, as in Fig. 8. *Left*: Observed correlation functions from Myers et al. (2007b, *gray circles*) and Hennawi et al. (2006, *gray diamonds*), measured for $\sim L_*$ quasars over the redshift ranges $z \sim 1$–2. Dashed black line shows the expected correlation function (nonlinear dark matter clustering from Smith et al. 2003 multiplied by the appropriate constant large-scale bias factor) without a small-scale excess. Gray lines multiply this by the predicted additional bias as a function of scale from § 2.3, namely, the fact that small-scale overdensities increase the probability of mergers. Solid line shows our mean prediction, dashed line the approximate $\sim 1\ \sigma$ range, as in Fig. 8. *Center*: Same, but dividing out the best-fit large-scale correlation function (i.e., bias as a function of scale). Black squares in upper panel show the measurement for true optical quasars ($-23.3 > M_i > -24.2$) from Serber et al. (2006) at $z \sim 0.1$–0.5. *Right*: Ratio of the mean bias at small radii ($r < 100\ h^{-1}$ kpc) to that at large radii (*the asymptotic values in the center panel*), at all redshifts where this has been observed. Lines show the predicted excess from the previous panels (lower line averages down to a minimum radius $r = 50\ h^{-1}$ kpc, upper line to a—potentially unphysical—minimum $r = 10\ h^{-1}$ kpc). [*See the electronic edition of the Supplement for a color version of this figure.*]

(McLure & Dunlop 2004; Kollmeier et al. 2006), active BH mass functions (Vestergaard 2004; Fine et al. 2006; Greene & Ho 2008), and cosmic background measurements (Volonteri et al. 2006; Hopkins et al. 2007e).

The quasar luminosity function $\phi(L)$ is given by the convolution over the merger rate (rate of formation of BHs of final mass M_{BH} in mergers) and quasar lifetime (differential time spent at luminosity L by a BH of final mass M_{BH}):

$$\phi(L) = \int t_Q(L|M_{BH})\dot{n}(M_{BH}|z)d\log M_{BH}. \quad (8)$$

Note this technically assumes $t_Q \ll t_H$, but this is true for all luminosities and redshifts of interest here. Figure 18 shows this prediction at a number of redshifts, compared to the large compilation of QLF measurements from Hopkins et al. (2007e). The agreement is surprisingly good at all redshifts. At the most extreme luminosities $L_{bol} > 3 \times 10^{14}\ L_\odot$ at each redshift, our predictions may begin to fall short of the observed QLF, but this somewhat expected, as these luminosities naively imply $>10^{10}\ M_\odot$ BHs accreting at the Eddington limit. It is therefore likely that a full resolution at the most extreme luminosities involves either revising the estimate of these bolometric luminosities (i.e., the bolometric corrections adopted may not be appropriate for the most extreme objects, or there may be beaming effects) or including processes beyond the scope of our current investigation (e.g., super-Eddington accretion or multiple mergers in massive BCGs). Nevertheless, our simple merger-driven scenario appears to accurately predict the distribution and evolution of most quasar activity.

Integrating the QLF over the appropriate range, we trivially obtain the active fraction and can calculate this separately for each host mass M_{gal} or BH mass M_{BH} in equation (8). Figure 19 compares this to observations at both low and high redshift, for systems with $\dot{m} \equiv L/L_{Edd} > 0.1$, representative of typical Seyfert and quasar populations (e.g., McLure & Dunlop 2004). Note that the quasar lifetime integrated above this threshold is close to a

constant value $\lesssim 10^8$ yr, similar to observational estimates (Martini 2004). At very low masses/levels of activity, other fueling mechanisms may be dominant—for comparison with, e.g., the active fractions in Hao et al. (2005) of typical $\lesssim 10^7\ M_\odot$ BHs ($\lesssim 10^{10}\ M_\odot$ hosts), we refer to secular and/or ''stochastic'' accretion models in disks (e.g., Hopkins & Hernquist 2006) and old ellipticals (Simões Lopes et al. 2007). Furthermore, at the lowest masses plotted, the typical AGN luminosities become extremely faint (typical $M_B \gtrsim -18$ in $M_{gal} \lesssim 10^{10}\ M_\odot$ hosts), and so such systems may be more often classified as non-AGN or typical star-forming systems (e.g., Rodighiero et al. 2007). At high levels of accretion, however, the merger-driven prediction agrees well with observations at low and high redshift and predicts a downsizing trend similar to that seen—namely, that from $z = 2$ to $z = 0$, quasar activity has been particularly suppressed in the most massive systems (although it has been suppressed to some extent at all host masses), presumably due to the conversion of these systems to ''red and dead'' spheroids without cold gas supplies (see Paper II).

We next follow Lidz et al. (2006) and extend equation (8) to convolve over the expected bias of the active systems at each quasar luminosity L,

$$b(L) = \frac{1}{\phi(L)} \int b(M_{BH})t_Q(L|M_{BH})\dot{n}(M_{BH}|z)d\log M_{BH}, \quad (9)$$

where $b(M_{BH})$ is determined just as $b(M_{gal})$ in § 2.4, by convolving over the contributions to each merging range in M_{BH} from all M_{halo}. Figure 20 plots the expected bias as a function of luminosity at each of several redshifts. As originally demonstrated in Lidz et al. (2006) our model for quasar light curves and the underlying triggering rate of quasars predicts a relatively weak dependence of clustering on quasar luminosity. Here we essentially rederive this result with an a priori prediction of these triggering rates, as opposed to the purely empirical (fitted to the QLF) rates from Lidz et al. (2006) and find that the conclusion is robust. However, this prediction is not necessarily a consequence
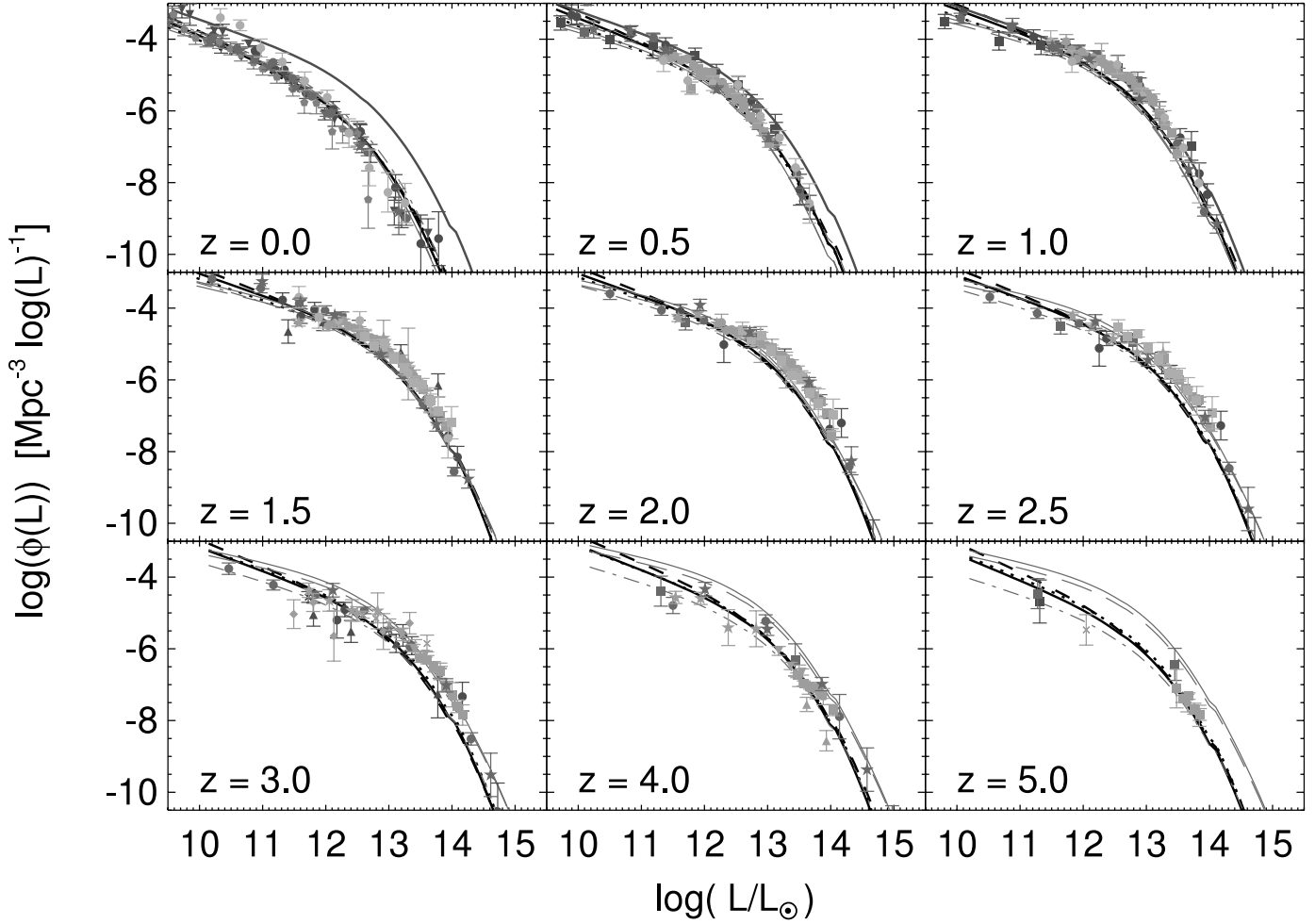
FIG. 18.—Predicted quasar luminosity functions, convolving our predicted merger rate functions (Fig. 9; same line styles) with quasar light curves from simulations (Hopkins et al. 2006a). Red lines allow dry mergers to trigger quasar activity as well (leading to an overestimate at low redshifts, as in Fig. 13). Points show observed bolometric luminosity functions at each redshift, from the compilation of observations in Hopkins et al. (2007e). QLF measurements derived from observations in the optical, soft X-ray, hard X-ray, mid-IR, and narrow emission lines are shown, respectively. The merger-driven model naturally predicts the observed shape and evolution of the QLF at all redshifts. [*See the electronic edition of the Supplement for a color version of this figure.*]

of merger-driven models (nor is it unique to them)—we show the predictions from the semianalytic models of Wyithe & Loeb (2002) and Kauffmann & Haehnelt (2000), who adopt simplified "lightbulb"-like quasar light curves (for a detailed discussion of these differences, see Hopkins et al. 2007d).

The reason for the weak dependence of quasar clustering on luminosity in Figure 20 is, in fact, the nature of the quasar light curve. Quasars grow rapidly in mergers to a peak quasar phase at the final stages of the merger, which exhausts and expels the remaining gas, after which the quasar decays to lower luminosities. This decay moves objects of the same host properties to fainter luminosities in the QLF, making the clustering properties flat as a function of luminosity. Thus, while an important test of our modeling (that the correct halos and galaxies host quasars of the appropriate luminosities), this is not a unique prediction of merger-driven models.

We can also use our model to estimate the infrared luminosity functions of various populations versus redshift. By construction, our assumed halo occupation model reproduces the observed star-forming (blue) galaxy mass function at each redshift. Using the corresponding fitted star formation histories as a function of baryonic mass from Noeske et al. (2007) (which fit the observations locally and their evolution at least to $z \sim 1.5$), we immediately obtain an estimate of the star formation rate function in "quies-

cent" (nonmerging) galaxies at each redshift. We include a scatter of ~ 0.25 dex in SFR at fixed stellar mass, comparable to that observed (in blue galaxies), but this makes relatively little difference, as the most extreme SFR populations are dominated by mergers. We then adopt the standard conversion from Kennicutt (1998) to transform this to an infrared luminosity function (where we refer to the total IR 8–1000 μm luminosity).

Our model also yields the mass function of gas-rich mergers, for which we can estimate their distribution of star formation rates. In Hopkins et al. (2006f) we quantify the distribution of star formation rates as a function of galaxy properties from the same large suite of simulations used to estimate the quasar lifetime. Essentially, this quantifies the "lifetime" above a given SFR in a merger, which can be reasonably approximated as a simple function of galaxy mass and (premerger) gas fraction:

$$t(>\dot{M}_*) = t_* \Gamma\left(0, \frac{\dot{M}_*}{M_f f_{\mathrm{gas}}/t_*}\right), \qquad (10)$$

where M_f is the postmerger galaxy mass (i.e., our M_{gal}) and $t_* \approx 0.3$ Gyr is a fitted characteristic time. This functional form simply amounts to the statement that there is a mean characteristic timescale t_* in which most of the gas mass of the merger ($M_f f_{\mathrm{gas}}$) is converted into stars, which we find is (unsurprisingly) similar
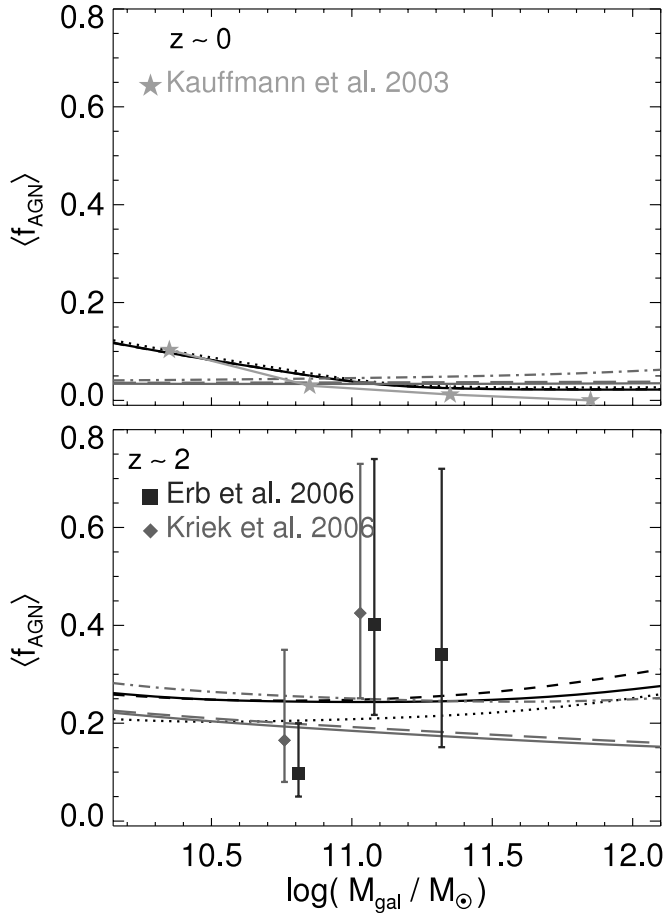
Fig. 19.—Predicted AGN fraction as a function of host properties. *Top:* Low-redshift quasar fraction (defined here by Eddington ratios $\dot{m} > 0.1$) as a function of galaxy mass. Black lines show the prediction of our merger-driven model, in the style of Fig. 9. Observed fractions are shown down to (roughly) their completeness limit, from Kauffmann et al. (2003). *Bottom:* Same, but at $z \approx 2$, with the AGN fraction determined observationally in LBG (Erb et al. 2006) and K-selected (Kriek et al. 2007) samples. Some caution should be applied at $M_{\rm gal} \lesssim 10^{10}\ M_{\odot}$, as the AGN luminosities become sufficiently low that even moderate star formation will dominate the observed luminosity and systems may not be classified as AGNs. [*See the electronic edition of the Supplement for a color version of this figure.*]

to the dynamical time of the merger and to observational estimates of the characteristic star formation timescale in starbursts and ULIRGs (Kennicutt 1998). Since the fitted star formation histories of Noeske et al. (2007) implicitly define a gas fraction as a function of time (or can be used in combination with the Schmidt-Kennicutt star formation law to infer the gas fraction), we simply adopt these for the premerger galaxies (but we have checked that they correctly reproduce observed gas fractions as a function of mass at $z = 0, 1, 2$; see Hopkins et al. 2007c). It is worth noting that, with this estimate, the explicit dependence on $f_{\rm gas}$ can be completely factored out in equation (10), and we can write it as an estimate of the amount of time a system spends above a given enhancement in SFR (basically a merger enhances the τ-model SFR by $\sim \tau/t_*$), relative to the premerger SFR. Using the same SFR to $L_{\rm IR}$ conversion, we obtain a rough estimate of the IR luminosity function of mergers.

Finally, adopting the empirically calculated obscured fraction as a function of quasar luminosity from Gilli et al. (2007) and assuming that the obscured bolometric luminosity is reradiated in the IR, we convert our predicted bolometric QLF to an IR QLF of obscured quasars. Technically, not all of the luminosity will be obscured, of course, but we find that, e.g., using the full dis-

tribution of column densities as a function of quasar luminosity from Ueda et al. (2003) to attenuate a template AGN SED yields a very similar answer (see also Franceschini et al. 2005), as does using a mean X-ray to IR bolometric correction of obscured AGNs (Elvis et al. 1994; Zakamska et al. 2004; Polletta et al. 2006). Including the IR contribution from unobscured quasars is a negligible correction.

Figure 21 compares the resulting predicted IR luminosity functions to observations at $z = 0$–2 and to the observed IR luminosity density at $z \sim 0$–5. At all redshifts, the agreement is good, which suggests that our model accurately describes the star formation history of the universe. This should be guaranteed, since at all redshifts the quiescent population dominates the $\sim L_*$ optical and IR luminosity functions (hence also the star formation rate and IR luminosity densities)—at this level, we simply confirm that our halo occupation model is a good approximation. However, at high luminosities, typical of ULIRGs, the populations are generally dominated by mergers and (at the highest luminosities) obscured AGNs.

We explicitly quantify the transition point as a function of redshift in Figure 22 (we show the comparison there just for our "default" model, but as is clear in Fig. 21, the transition between different populations dominating the LF is similar regardless of the exact version of our model adopted). Our comparisons generally affirm the conventional wisdom: at low redshift, mergers dominate the ULIRG and much of the LIRG populations, above a luminosity $\sim 10^{11.4}\ L_{\odot}$, with heavily obscured (potentially Compton-thick) AGNs (in starburst nuclei) becoming a substantial contributor to IR-luminous populations in the most extreme $\gtrsim$ few $\times\ 10^{12}\ L_{\odot}$ systems (nearing hyper-LIRG $>10^{13}\ L_{\odot}$ luminosities, which are common bolometric luminosities for $>10^8\ M_{\odot}$ BHs near Eddington, but would imply potentially unphysical $\gtrsim 1000\ M_{\odot}\ \mathrm{yr}^{-1}$ SFRs). At higher redshifts, disks are more gas-rich and thus have characteristically larger star formation rates, dominating the IR LFs at higher luminosities. By $z \sim 1$, most LIRGs are quiescent systems, and by $z \sim 2$, only extreme systems $\gtrsim$ few $\times\ 10^{12}\ L_{\odot}$ are predominantly mergers/AGNs.

This appears to agree well with recent estimates of the transition between AGNs and passive star formation dominating the bolometric luminosities of high-redshift systems. Interestingly, this shift occurs even while increasing merger rates (and higher gas fractions in typical mergers) lead to a larger overall contribution of mergers to the star formation rate and IR luminosity densities. At $z \sim 0$, mergers contribute negligibly to the total IR luminosity density, but by $z \sim 2$, they may contribute $\sim 20\%$–50% of the IR output of the universe, with that contribution due comparably to both star formation in mergers and obscured BH growth (which should be true, given the $M_{\rm BH}$-$M_{\rm host}$ correlations and typical $\epsilon_r \sim 0.1$ radiative efficiencies; see, e.g., Lidz et al. 2007).

The integrated contribution of mergers to the star formation rate and IR luminosity densities agrees well with observational estimates (available at $z \lesssim 2$; see Bell et al. 2005; Menanteau et al. 2006), and the constraint from stellar population models that only a small fraction of the $z = 0$ stellar mass in typical early-type galaxies was formed in the spheroid-forming merger itself (as opposed to more extended star formation in the premerger disks; e.g., Noeske et al. 2007). For a more detailed comparison and analysis of the merger-induced contribution to the star formation rate density of the universe, we refer the reader to Hopkins et al. (2006f).

We caution that the above comparisons are approximate and intended as a broad check that our models are consistent with the observed abundance of IR luminous galaxies as a function of
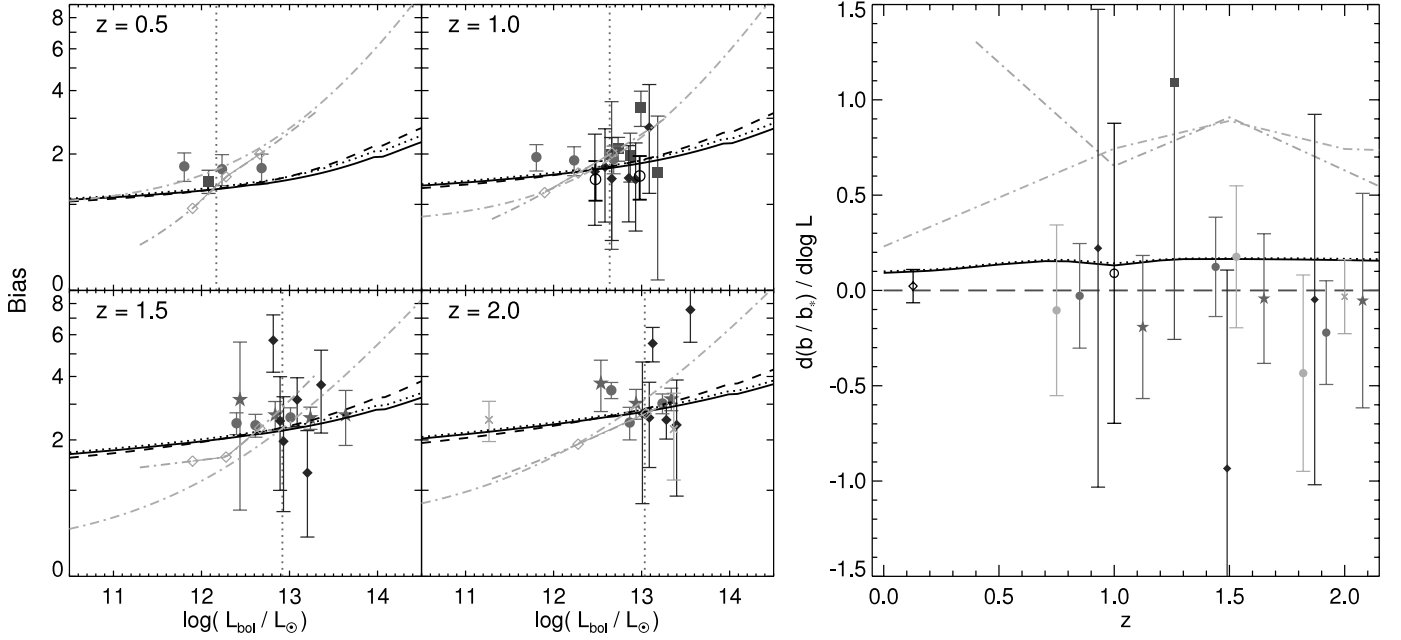
Fig. 20.—*Left*: Predicted bias as a function of quasar luminosity from our merger-driven model (*black lines*, style as in Fig. 9). To contrast, the expected bias $b(L)$ from the semianalytic models of Wyithe & Loeb (2002, *light gray*) and Kauffmann & Haehnelt (2000, *light gray with diamonds*) are plotted (*dot-dashed lines*); these adopt simplified (constant or exponential "on/off") quasar light curves. Points are measurements from Croom et al. (2005, *gray squares*), Adelberger & Steidel (2005, *light gray crosses*), Porciani & Norberg (2006, *dark gray diamonds*), Myers et al. (2007a, *gray circles*), da Angela et al. (2007, *gray stars*), and Coil et al. (2007, *black open circles*). For ease of comparison, all luminosities are converted to bolometric luminosities using the corrections from Hopkins et al. (2007e). Vertical gray dotted lines show L_* in the QLF at each redshift, from Hopkins et al. (2007e). *Right*: The best-fit slope of the dependence of bias on luminosity at the QLF L_*, i.e., $d(b/b_*)/d\log(L/L_*)$, where $b_* \equiv b(L_*)$. Points are determined from the observations at left, with the observations from Myers et al. (2007a, *light gray circles*) and Grazian et al. (2004) and Wake et al. (2004, *black open diamond*) added. Lines are in the style of the left panel, with the gray dashed line showing no dependence of bias on luminosity. Adopting an a priori model for merger-triggered quasar activity reproduces the empirical prediction from Lidz et al. (2006) that quasar bias should depend weakly on quasar luminosity. [*See the electronic edition of the Supplement for a color version of this figure.*]
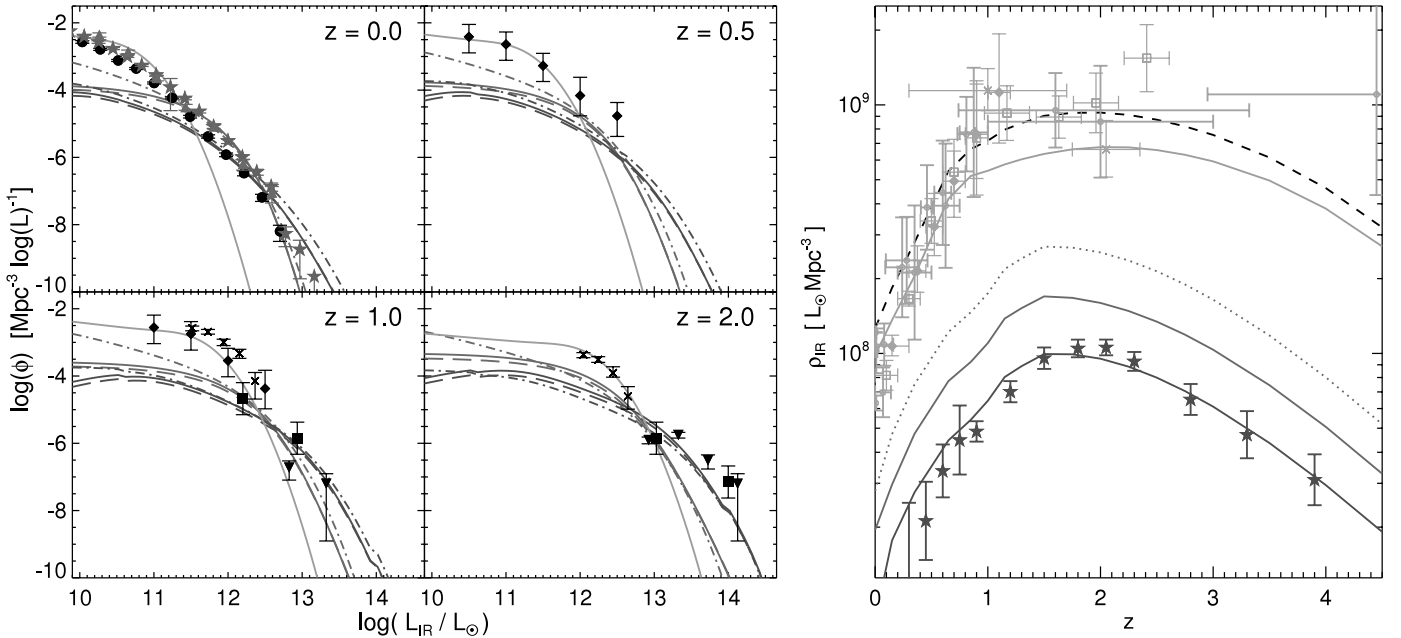

Fig. 21.—*Left*: Predicted total IR (8–1000 μm) luminosity functions at different redshifts (as labeled). Gray lines shows the estimated contribution from nonmerging systems, star formation in mergers, and obscured AGN in mergers. Line styles are as in Fig. 9, for the variants of the merger calculations. Points show observational estimates from Saunders et al. (1990, *gray stars*), Soifer & Neugebauer (1991, *gray triangles*), Yun et al. (2001, *black circles*), Le Floc'h et al. (2005, *black diamonds*), Chapman et al. (2005, *black inverted triangles*), Babbedge et al. (2006, *black squares*), and Caputi et al. (2007, *black crosses*. *Right*: Integrated IR luminosity density. Solid lines show the contributions from nonmerging systems (*gray*), star formation in mergers (*gray*), and obscured quasars in mergers (*gray*). Gray dotted line shows the total (star formation+AGN) merger contribution, black dashed line shows the total from all sources. Gray points show observational estimates of ρ_{IR} from the compilation of Hopkins (2004, *circles*; only the direct IR observations therein are plotted here, as well as Le Floc'h et al. (2005, *diamonds*), Pérez-González et al. (2005) and Caputi et al. (2007, *crosses*). Gray stars show the bolometric quasar luminosity density from Fig. 13, rescaled by a constant (mean) obscured-to-unobscured ratio of $\sim$2:1. The agreement in all cases is good—our model reproduces the star formation history of the universe and distribution of star formation rates and bolometric luminosities. [*See the electronic edition of the Supplement for a color version of this figure.*]
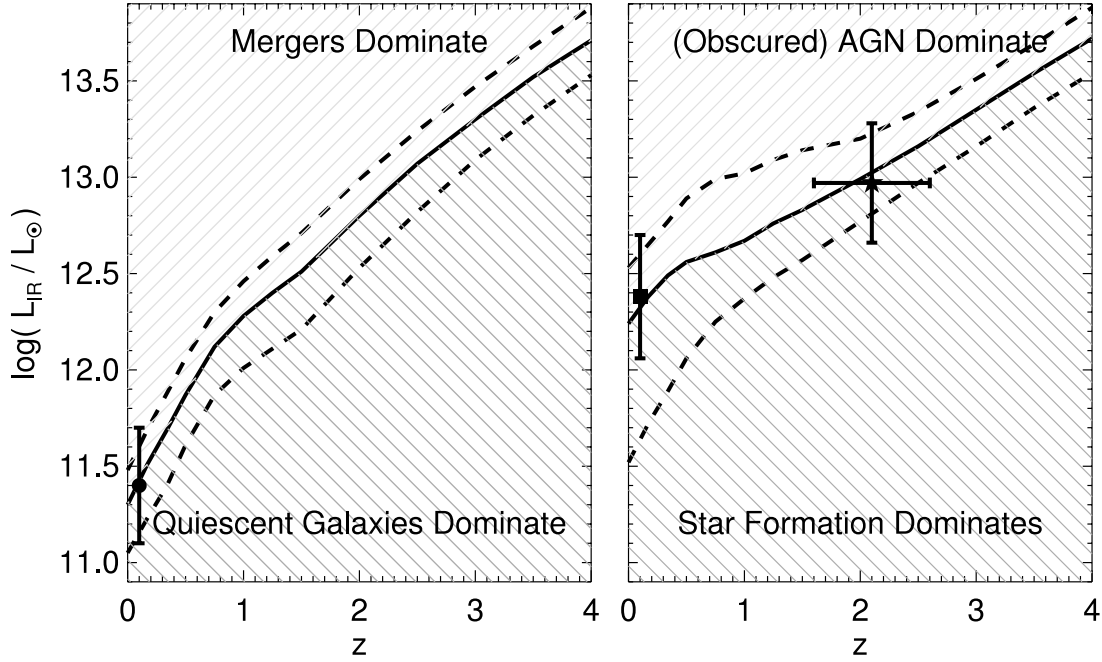
Fig. 22.—*Left*: Total IR luminosity, as a function of redshift, above which mergers (star formation+AGN) dominate the total IR luminosity functions (*solid line*, from Fig. 21; dashed lines show the range above which 25%/75% of systems on the luminosity function are mergers). Point shows the corresponding transition point (and range) observed in low-redshift systems (Sanders et al. 1988b). *Right*: Same, but for the transition between star formation (in nonmerging+merging systems) and (obscured) AGNs dominating the IR luminosity functions (generally a factor of ∼few larger luminosity than the quiescent system-merger transition). Points show the observed estimates from comparison of PAH feature strengths in Lutz et al. (1998, low redshift) and Sajina et al. (2007, high redshift). A similar estimate is obtained (at low redshift) from comparison of emission line strengths (Sanders & Mirabel 1996; L. Kewley et al. 2008, in preparation), full SED template fitting (Farrah et al. 2003), or indirect comparison with type 2 AGN luminosity functions (Chary & Elbaz 2001). The model predicts the local transitions, and that by $z \gtrsim 1$, the LIRG population is dominated by quiescent star formation in gas-rich systems (even as the total and fractional luminosity density in mergers increases rapidly). [*See the electronic edition of the Supplement for a color version of this figure.*]

redshift. We have ignored a number of potentially important effects: for example, obscuration is a strong function of time in a merger and may affect various luminosities and morphological stages differently. Moreover, our simple linear addition of the star formation contribution of mergers to the IR LF and the AGN contribution is only technically correct if one or the other dominates the IR luminosity at a given time in the merger; however, there are clearly times during the final merger stages when the contributions are comparable. Resolving these issues requires detailed, time-dependent radiative transfer solutions through high-resolution simulations that properly sample the merger and quiescent galaxy parameter space at each redshift and is outside the scope of this work (although an important subject for future, more detailed study; see, e.g., Li et al. 2008). It would be a mistake, therefore, to read too much into, e.g., the detailed predictions for submillimeter galaxies or other extreme populations based on Figures 21 and 22. However, most of our predicted qualitative trends, including the evolution of the luminosity density (and approximate relative contribution of mergers) and the shift in which quiescent or merger-driven populations dominate the bright IR LF, should be robust. Critically, a model in which merger-driven quasar activity dominates the QLF predicts an abundance of IR-luminous galaxies consistent with the observations as a function of both luminosity and redshift.

3.3. *When Merger-Triggering Loses to Secular Processes*

Despite these arguments for a merger-driven origin for bright, high-redshift quasars, there are good reasons to believe that most local, high Eddington ratio objects are *not* related to mergers. Most active local systems typically involve relatively low-mass BHs ($M_{\rm BH} \sim 10^7 \, M_\odot$; Heckman et al. 2004), in Sa/b-type host

galaxies, without significant evidence for recent major interactions (Kauffmann et al. 2003; Pierce et al. 2007), and have relatively low Seyfert-level luminosities ($-21 \gtrsim M_B \gtrsim -23$; Hao et al. 2005), below the traditional $M_B = -23$ Seyfert-quasar divide. Given this, it is natural to ask whether there are additional reasons to believe that bright quasars have distinct origins, and if so, when (or at what luminosities) these nonmerger-driven fueling mechanisms begin to dominate AGN populations.

In addition to the arguments in §§ 3.1 and 3.2, there are a number of qualitative differences between bright, high-redshift quasars and local Seyferts. Quasars have significantly different clustering amplitudes (Hopkins et al. 2007d) and host stellar mass distributions (Hopkins et al. 2007a) from star-forming galaxies at $z \gtrsim 1$, and they typically have hosts with elliptical or merger remnant morphologies (Floyd et al. 2004; Falomo et al. 2004; Zakamska et al. 2006; Letawe et al. 2007), frequently exhibiting evidence of tidal disturbances (Bahcall et al. 1997; Canalizo & Stockton 2001; Hutchings et al. 2003, 2006; Urrutia et al. 2008; Bennert et al. 2008). Figure 23 compares the clustering as a function of scale measured in Serber et al. (2006) for both bright quasars and Seyfert galaxies—quasars exhibit the strong trend of excess clustering on small scales indicative of a triggering process that prefers small-scale overdensities, but Seyferts show no significant preference for local overdensities.

Because galaxy mergers are also associated with the termination of star formation in the remnant (even if only temporarily), i.e., a rapid poststarburst phase and transition to the red sequence (discussed in detail in Paper II), the decay of the quasar light curve should be associated with the reddening of the remnant, in a merger-driven model. This implies a particular preferred track for quasar hosts in the color-magnitude diagram, illustrated in
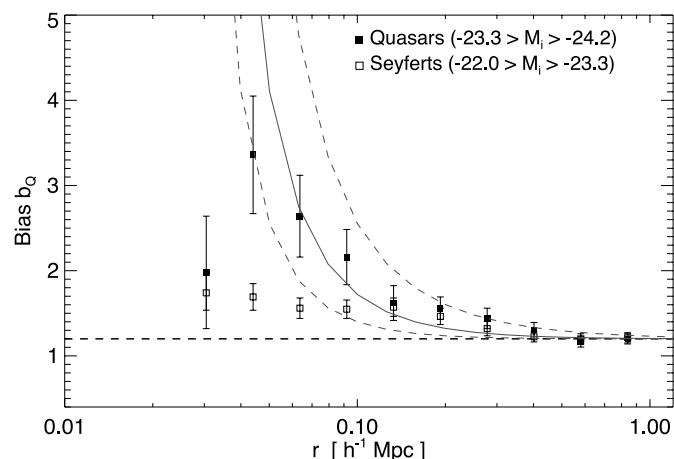
FIG. 23.—Same as Fig. 17 (*upper center panel*), but comparing the clustering (quasar-galaxy cross-correlation) as a function of scale measured by Serber et al. (2006) for bright optical quasars and dimmer Seyfert galaxies. Quasar clustering is consistent with our predicted excess on small scales, indicating a merger-driven origin, but low-luminosity systems show no such dependence, suggesting that processes independent of the local, small-scale density (e.g., secular processes) may dominate at these luminosities. [*See the electronic edition of the Supplement for a color version of this figure.*]

Figure 24. In this scenario, quasars should be associated with the crossing of the "green valley"—i.e., the triggering of a quasar occurs at the end of the merger, when young stellar populations imply a bluer-than-average host spheroid, and the quasar decays to lower luminosities as the remnant reddens onto the red sequence.

Alternatively, if quasars were triggered in a purely secular manner, or otherwise independent of whatever quenching mechanism terminates the galactic supply of cold gas, then their natural preferred location is in the blue cloud—i.e., blueward of the "green valley." Systems in this regime still have cold gas supplies and have not yet quenched. Because the quenching is uncorrelated with quasar triggering in such a model, and the lack of galaxies in the "green valley" implies that this transition is rapid, very few quasars would be expected to be triggered just as the quenching occurs, and therefore few quasars should be present in the "green valley."

Comparing these qualitative scenarios with observations appears to favor the former, merger-driven case. Quasars tend to live redward of the "top" of the blue cloud, with the brightest/highest accretion rate quasars preferentially in bluer-than-average spheroids in the "green valley" (Kauffmann et al. 2003; Sánchez et al. 2004; Nandra et al. 2007).

Figure 25 shows this quantitatively—we plot the distribution of colors of quasar hosts, compared with that fitted to the blue cloud and red sequence, or systems with observed bars and/or disk instabilities (the expected quasar hosts in a secular model, regardless of quasar duty cycles during a bar phase), and poststarburst (E+A/K+A) systems, largely identified as merger remnants and "blue spheroids" (see the discussion in § 2.3). The quasar hosts clearly lie preferentially between the blue cloud and red sequence, with a color distribution very similar to observed poststarburst galaxies.

The distribution is quite distinct, however, from observed barred systems, which lie overwhelmingly on the blue sequence with, if anything, a bias toward the bluest systems (which is expected, as these are the most gas-rich and therefore most unstable systems). Even if one assumes that, in the most extreme bar instabilities, dust reddening might move the system into the "green valley" as a reddened disk, this appears to contradict the observations above,
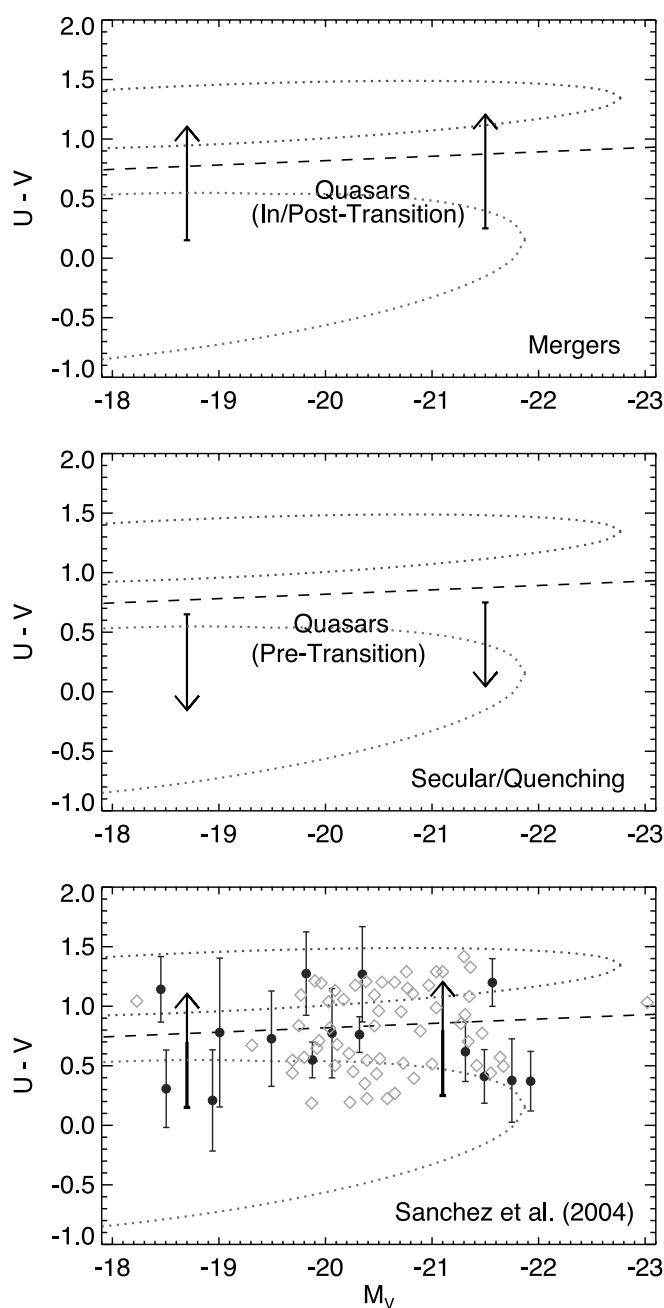

FIG. 24.—Location of quasars in the color-magnitude diagram, expected from different models. *Top*: The top and bottom dotted regions roughly outline the red sequence and blue cloud, respectively, with the dashed line dividing the bimodality (from Bell et al. 2004). Arrows show the preferred location of quasar hosts in a merger-driven model. At the end of a merger, a bright quasar is triggered in a spheroid/merger remnant at the top of the blue cloud (due to the young stellar populations from premerger and merger-induced star formation), and subsequently the quasar luminosity decays while the remnant rapidly reddens, leaving a relatively low accretion rate remnant on the red sequence. *Middle*: Same, but for a secular triggering scenario in which quasar activity (which must still require cold gas) is uncorrelated with quenching or itself exhausts the gas supply. In this case, quasars should live in the blue cloud, with gas-rich systems, and their abundance rapidly drops approaching the "green valley" as gas supplies are exhausted. *Bottom*: We compare with observations of quasar host galaxy colors at $z \sim 0.7$–1.1 from Sánchez et al. (2004, circles). X-ray identified AGN and quasar hosts from Nandra et al. (2007, diamonds) are also shown (the numbers plotted should not be taken literally, as we have rescaled the authors $U - B$ vs. M_B color-magnitude relation to that shown here for the sake of direct comparison, but the result is qualitatively identical to that shown). Arrows reproduce the merger expectation from the top panel. Quasars appear to live in the region of color-magnitude space expected if they are triggered at the *termination* of star formation and subsequently decay in luminosity, as expected in merger-driven scenarios. [*See the electronic edition of the Supplement for a color version of this figure.*]
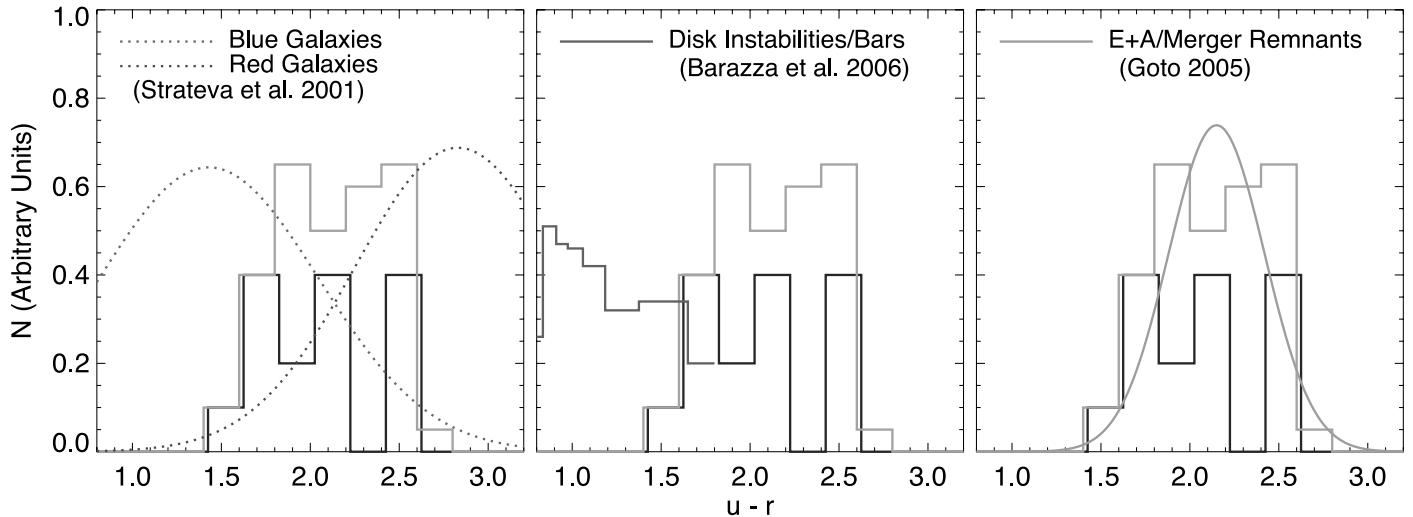
FIG. 25.—Distribution of quasar host galaxy colors from Fig. 24 (*histograms*; from Sánchez et al. 2004 and Nandra et al. 2007 in black and light gray, respectively). We compare with fitted (Gaussian) color distributions of blue cloud and red sequence galaxies from Strateva et al. (2001) with the distribution of colors of barred galaxies in the SDSS from Barazza et al. (2007; the expected quasar hosts in a secular or instability-driven quasar fueling model), and with the fitted (Gaussian) distribution of post-starburst (generally merger remnant) E+A/K+A galaxies in Goto (2005). Quasar host colors follow the "transition" between blue cloud and red sequence observed and expected in merger remnants, in contrast to the preferentially most gas-rich, blue hosts of observed strong bars. [*See the electronic edition of the Supplement for a color version of this figure.*]

which find quasars to be in preferentially blue spheroids (even X-ray observations, which suffer less severe bias against dust-reddened systems). A more rigorous quantitative comparison of the tracks through color-magnitude space and the relative abundances in this transition region will be the topic of future work (S. Wuyts et al. 2008, in preparation), and we stress that these are all relatively low-redshift samples, but studying how the mean quasar luminosity and accretion rates scale/decay with the degree of reddening or aging of their host stellar populations can provide a powerful discriminant between these models.

There are a number of additional constraints we can place on the contribution to the QLF from secular fueling in nonmerging disks. Figure 26 considers several of these. First, we place a limit on secular activity by asking, at a given z, what are the brightest QSOs possible in disk/star-forming galaxies? For that redshift, we take the observed mass function of star-forming galaxies and convolve with $P(M_{BH}|M_{gal})$ to obtain the hosted BH mass function (assuming the most massive disks are Sa/b-type galaxies). Then, assume that every such BH is at its Eddington luminosity. At some point (corresponding to $\gtrsim 2$–$4\,M_*$ in the disk mass function) the number density of these mock quasars falls below the QLF (which declines much less rapidly) at that luminosity and redshift. In other words, at high luminosities, the required BH masses from the Eddington limit are too large to live in late-type galaxies. To be optimistic, we assume *all* the quasar luminosity density below this limit is contributed by secular activity in disks. This then gives an upper limit to the fraction of the luminosity density from disks. We repeat this procedure for a number of different mass functions at different redshifts. In all cases, even this limit falls to a fraction $\ll 1$ by $z \gtrsim 1$, as the QLF L_* reaches large luminosities corresponding to $M_{BH} \gtrsim 10^8\,M_\odot$ BHs at the Eddington limit. Given the BH-host spheroid mass relations, this requires a very massive spheroid, easily formed in a merger, but not present in even the most early-type disks.

Second (alternatively), we assume all BHs in pseudobulges were formed via secular mechanisms. As discussed in § 1, there is good reason to believe that this is the case, whereas classical bulges must be formed in mergers. For a given $z = 0$ BH population, we infer an accretion history in the standard fashion from

matching the BH mass function and continuity equations (e.g., Salucci et al. 1999; Yu & Tremaine 2002). We then calculate the fraction of the QLF luminosity density at a given redshift from systems that, at $z = 0$, live in pseudobulges. We consider this for several different observational estimates of the pseudobulge fraction as a function of, e.g., host galaxy morphological type or bulge Sersic index (Kormendy & Kennicutt 2004; Balcells et al. 2007; Allen et al. 2006; Noordermeer & van der Hulst 2007) and the directly estimated pseudobulge mass functions in Driver et al. (2007). Although the details are sensitive to how we define pseudo-bulges, we find a similar result—massive BHs that dominate the luminosity density at $z \gtrsim 1$ live in the most massive bulges/ellipticals, which are overwhelmingly classical bulges.

Third, we calculate the probability that the observed clustering of quasars is consistent with that of star-forming/disk galaxies (see Fig. 15). This is subject to some important caveats—although quasar clustering depends only weakly on luminosity (see Fig. 20), galaxy clustering has been shown to depend quite strongly on galaxy luminosity/stellar mass (Norberg et al. 2002). We use the compilation of clustering data from Hopkins et al. (2007d) as in Figure 15. At $z \lesssim 1.5$, we specifically compare the clustering of $\sim L_*$ quasars with that of $\sim L_*$ blue/star-forming galaxies. For *any* model in which quasars are driven by secular activity and the statistics of quasar light curves/triggering are continuous as a function of host mass/luminosity (i.e., there is not a second feature in the luminosity function introduced by the statistics of the light curves themselves), these should roughly correspond. At higher redshift, galaxy clustering as a function of type and luminosity/mass at $\sim L_*$ is not clearly resolved, so we can only plot combined clustering of observed star-forming populations (generally selected as Lyman-break galaxies); again caution is warranted given the known dependence of clustering on galaxy mass/luminosity (for LBGs, see Allen et al. 2005). Fortunately, the range of particular interest here is $z \lesssim 1$, where we again find a similar trend—quasar clustering is consistent with secular fueling at $z \sim 0$, but by $z \sim 1$ this is no longer true. As discussed in Hopkins et al. (2007d) this appears to be contrary to some previous claims (e.g., Adelberger & Steidel 2005); however, in most cases where quasars have been seen to cluster similarly to blue
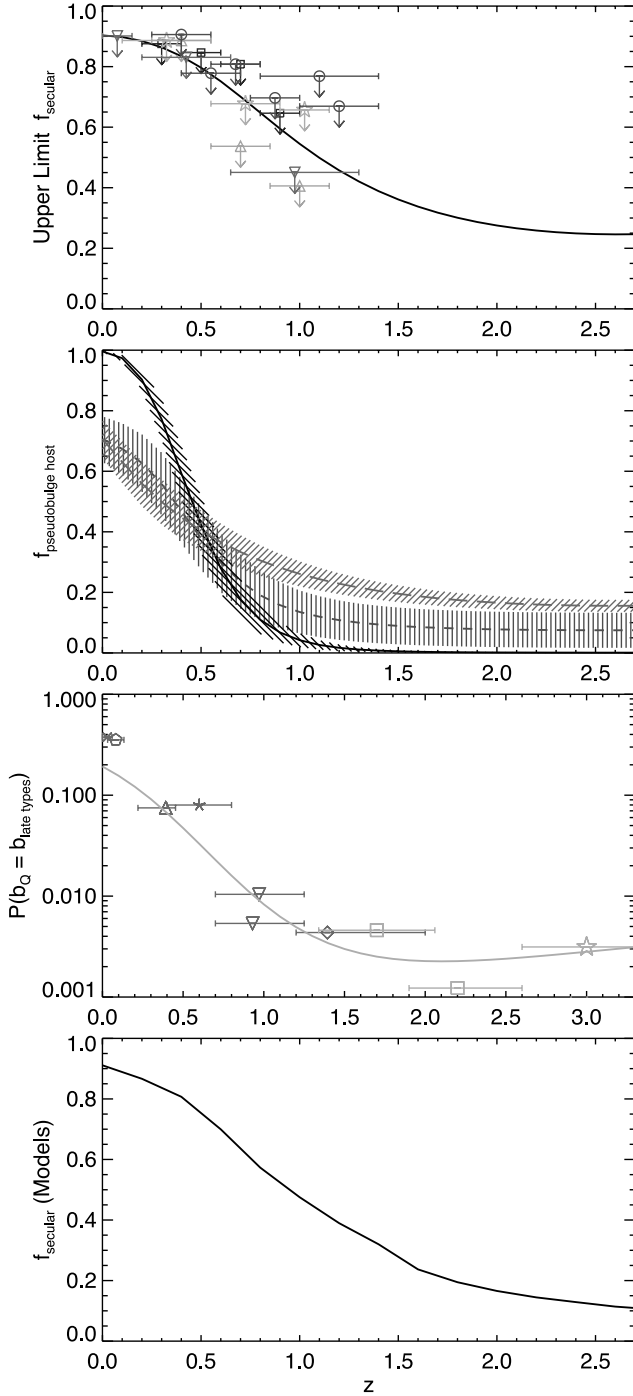
FIG. 26.—Fraction of the integrated quasar luminosity density due to nonmerger-driven secular mechanisms. *Top*: Upper limit to the contribution from BHs in disk galaxy hosts at each z (see text). Limits are derived from the observed type-separated mass functions in Fig. 16 (same style) and Franceschini et al. (2006, *cyan stars*). Solid line assumes the disk mass function does not evolve with z. *Second from top*: Fractional contribution from systems in pseudobulges at $z = 0$. Local distribution of pseudobulge masses is estimated from the observed pseudobulge fraction versus galaxy type (Noordermeer & van der Hulst 2007, *gray dashed line*, with ~1 σ shaded range), or assuming all bulges with Sersic index $n < 2$ are pseudobulges (with the distribution of n versus bulge mass from Balcells et al. 2007, *black solid line and shading*), or from directly measured pseudobulge mass functions (Driver et al. 2007, *gray long-dashed line and shading*). *Second from bottom*: Probability (from χ^2) that observed clustering of quasars (data in Fig. 15) and star-forming galaxies reflect the same hosts. Solid line is derived from the best-fit to the compilation of Hopkins et al. (2007d) points from the individual measurements included (see Fig. 15). *Bottom*: Predicted fraction of the luminosity density from the model for secular fueling from Hopkins & Hernquist (2006) when combined with the merger-driven model herein. [*See the electronic edition of the Supplement for a color version of this figure.*]

galaxies, either *faint* AGN populations (not $\sim L_*$ quasars) or bright ($\gg L_*$) blue galaxies were considered. Indeed, quasars do cluster in a manner similar to the *brightest* blue galaxies observed at several redshifts (e.g., Coil et al. 2007; Allen et al. 2005 at $z \sim 1$ and $z \gtrsim 2$, respectively). This should not be surprising; since quasars require some cold gas supply for their fueling, they cannot be significantly more clustered than the most highly clustered (most luminous) population of galaxies with that cold gas.

Finally, we compare these with a simple model expectation. We combine our prediction of the merger-driven QLF with the model from Hopkins & Hernquist (2006) for the QLF driven by secular fueling mechanisms in star-forming galaxies. This prediction is based on a simple model of feedback-driven self-regulation, calculating the rate of triggering in nonmerging disks from the observed statistics of gas properties in the central regions of star-forming galaxies of different types. The result is similar to the empirical constraints.

All of these comparisons have important caveats. For example, secular mechanisms could act so quickly as to completely transform disks to bulges, rapidly making very large BHs (although this conflicts with the pseudobulge constraints) from disk hosts. Pseudobulges could form in more systems than we estimated but be subsequently transformed to classical bulges via major mergers. Clustering could be affected by a number of systematic uncertainties inherent in, e.g., the mass and luminosity ranges considered. However, these systematics are independent, and there is no single loophole that can simultaneously reconcile the three constraints considered here with the possibility that secular fueling dominates bright $\sim L_*$ quasar activity at $z \gtrsim 1$. Although there are differences in detail, all the methods we have considered empirically suggest a similar scenario: secular (nonmajor merger related) fueling mechanisms contribute little to quasar activity at $z \gtrsim 1$, which involves the most massive $M_{\rm BH} \gtrsim 10^8 \ M_\odot$ BHs in the most massive spheroids. By $z \sim 0.5$, however, the most massive BHs are no longer active, and a significant fraction of the quasar luminosity density can come from $\sim 10^7 \ M_\odot$ BHs in undisturbed hosts. By $z \sim 0$, the local QLF is largely dominated by Seyfert activity in relatively small BHs with late-type, undisturbed host disks (Heckman et al. 2004).

Even if we ignore these constraints, a model in which secular fueling dominates the growth of quasars and BHs has difficulty matching the observed rise and fall of the quasar luminosity density with cosmic time. Figure 27 illustrates this. We show the observed bolometric quasar luminosity density as a function of redshift, compared to our estimate of the merger-driven luminosity density (as in Fig. 13). We also show our estimate of the luminosity density, which comes from systems that , at $z = 0$, live in pseudobulges, calculated as in Figure 26. Again, this fairly moderate, empirical model of secular activity can account for the observed luminosity density at low redshifts $z \lesssim 0.5$ but provides only a small contribution at high redshifts $z \gtrsim 1$.

We might, however, imagine a "maximal" secular model in which *all* spheroids are initially formed by disk instabilities. Equivalently (for our purposes), albeit highly contrived, a model might invoke secular processes to rapidly build up BH mass (to the final mass that will be given by the "future" $M_{\rm BH}$-σ relation) before a spheroid is formed in later mergers and/or instabilities. These have severe difficulty reconciling with the kinematics of observed classical bulges (see § 1) and the tightness of the BH-host spheroid correlations, respectively, and they are not favored by simple dynamical arguments (see, e.g., Shen et al. 2003), nor the constraints in Figure 26, but they could in principle be invoked. In fact, the semianalytic model of Bower et al. (2006) is effectively such a scenario, in which a very strong disk instability
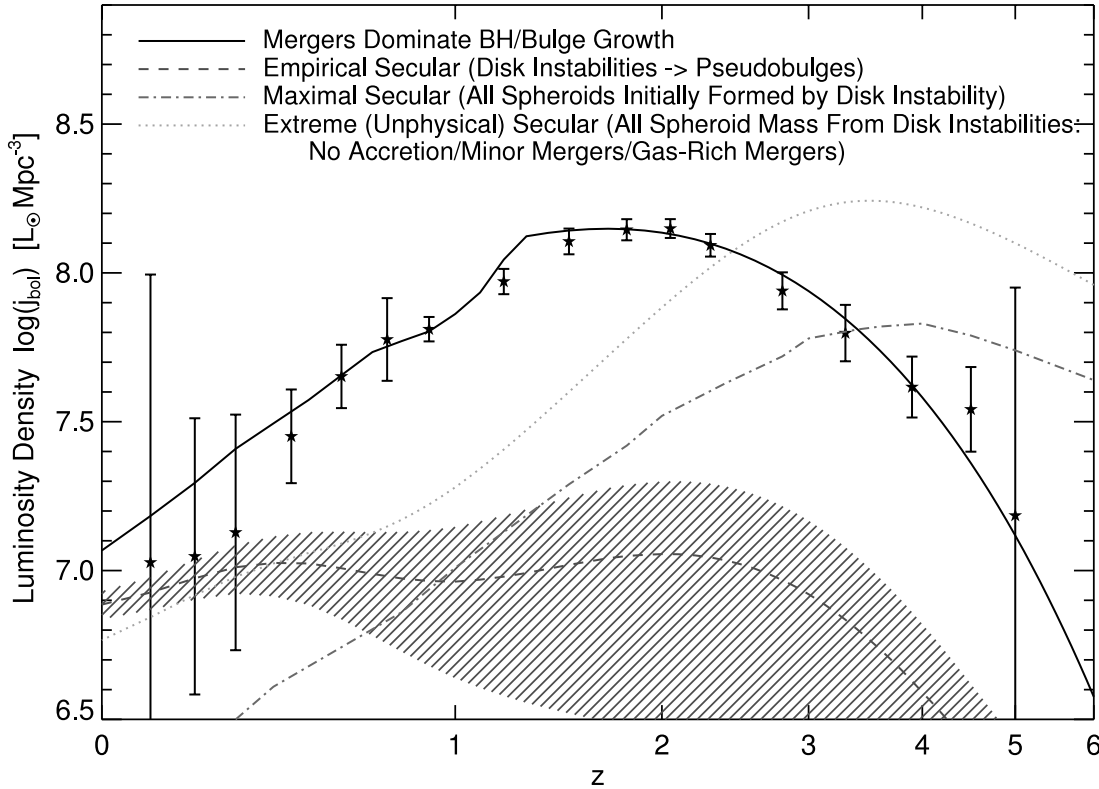
FIG. 27.—Bolometric quasar luminosity density as a function of redshift. Black stars show the observations from Hopkins et al. (2007e). Lines show estimates from different models (as labeled): the prediction from a merger-driven model (as in Fig. 13) and a moderate secular model in which BHs in pseudobulges at $z = 0$ were formed in disk instabilities (as in Fig. 26, line in same style) are in good agreement with the luminosity density evolution and empirical constraints on clustering, host galaxy colors, spheroid kinematics, and disk/spheroid mass functions. We compare a maximal secular model, from Bower et al. (2006) in which most BHs and (even classical) spheroids are initially formed via disk instabilities, and an "extreme" secular model, in which all $z = 0$ BH mass is formed in such instabilities (same as the maximal secular model, but with no BH growth from cooling, accretion, or mergers; this is unphysical but serves as a strong upper limit). In order for disk instabilities to dominate BH growth, they must act very rapidly, before the (inevitable) major mergers can exhaust gas and form massive spheroids—this forces such models to predict a luminosity density history offset to earlier times (higher redshifts) compared to the merger-driven model, in disagreement with the observations. [See the electronic edition of the Supplement for a color version of this figure.]

mode is analytically adopted that overwhelmingly dominates initial bulge formation and BH growth (mergers contributing ≪1% at all redshifts). We therefore compare their estimate for the total quasar luminosity density (accretion rate density) as a function of time. Finally, in the default Bower et al. (2006) model, there is still some growth of BHs via accretion from the diffuse ISM, cooling, and mergers (major and minor). We therefore also adopt an even more extreme secular model, in which we reproduce the Bower et al. (2006) analysis with an even stronger disk instability mode—essentially renormalizing the model such that all $z = 0$ bulge mass was formed in this "secular" mode (i.e., we allow *no* subsequent growth via other mechanisms and demand that the observed integrated $z = 0$ BH mass density be matched by the integrated secular mode growth). This latter model is of course unphysical but yields a hard upper limit to secular-mode growth.

It is immediately clear that the "maximal" secular model predicts that the quasar luminosity density should peak at much higher redshifts $z \sim 4$ than the observed $z \sim 2$. In general, the rise and fall of the quasar luminosity density in such a model are offset to earlier times. The reason for this is simple: in a fully cosmological model, mergers are *inevitable*. And, whether or not most quasars are triggered by mergers, it is extremely difficult to contrive a major, gas-rich merger without BH accretion and spheroid formation, with most of the gas being consumed by star formation. The only way that a secular or disk instability model can dominate the integrated buildup of BH mass and quasar luminosity density is to "beat mergers to the finish," i.e., to gener-

ally operate early and rapidly enough such that the BHs have been largely formed, and gas already exhausted, by the time massive galaxies undergo their first major mergers. In such models, then, one is forced to predict that the quasar luminosity density peaks at very early times and has largely declined (i.e., most of the gas in massive systems has already been exhausted) by $z \sim 2$.

Finally, this relates to a more general point. The quasar luminosity density (and especially the number density of bright quasars corresponding to ≳$10^8 M_\odot$ BHs at high Eddington ratio; see Fan et al. 2004; Richards et al. 2006b) declines rapidly at $z \gtrsim 2$–3 [roughly as $\sim(1 + z)^4$–$(1 + z)^6$], compared to the global star formation rate density of the universe, which is relatively flat at these redshifts [declining as $\sim(1 + z)^0$–$(1 + z)^{1.5}$ from $z \sim 2$–6; Hopkins & Beacom 2006]. This has long been recognized and cited as a reason why quasars and BH growth cannot explain reionization at high redshifts (since, similar to the global star formation history, the UV background declines slowly at these redshifts). It further implies that BH growth (at least at the masses of interest for our predictions here) cannot generically trace star formation. This places strong constraints on secular models, as above, as well as models in which essentially all high-redshift star formation is in bulges or some sort of dissipational collapse (e.g., Granato et al. 2004; Lapi et al. 2006). Some process must delay the formation of massive BHs while allowing star and galaxy formation to proceed efficiently at high redshifts. A natural explanation is that massive BH formation requires major mergers. In our model, at high redshifts, low-mass galaxies can efficiently

form (and potentially build low-mass BHs via secular instabilities), but they are predominantly disks, which efficiently turn gas into stars and do not form very massive bulges or BHs. Only later, once their hosts have grown more massive, are they likely to undergo major mergers, which transform the disks into spheroids and build correspondingly massive BHs. This automatically explains the much sharper rise and fall of the quasar luminosity density and number density of bright quasars, relative to the shallow evolution in the star formation rate density and ionizing background of the universe at high redshifts.

4. DISCUSSION

We have developed a theoretical model for the cosmological role of galaxy mergers, which allows us to make predictions for various merger-related populations such as starbursts, quasars, and spheroidal galaxies. By combining theoretically well-constrained halo and subhalo mass functions as a function of redshift and environment with empirical halo occupation models, we can estimate where galaxies of given properties live at a given epoch. This allows us to calculate, in an a priori cosmological manner, where major galaxy-galaxy mergers occur and what kinds of galaxies merge, at all redshifts.

We compare these estimates to a number of observations, including observed merger mass functions; merger fractions as a function of galaxy mass, halo mass, and redshift; the mass flux/mass density in mergers; the large-scale clustering/bias of merger populations; and the small-scale environments of mergers, and we show that this approach yields robust predictions in good agreement with observations and can be extended to predict detailed properties of mergers at all masses and redshifts. There are some uncertainties in this approach. However, we recalculate all of our predictions adopting different estimates for the subhalo mass functions and halo occupation model (and its redshift evolution) and find this makes little difference (a factor <2) at all redshifts. The largest uncertainty comes from our calculation of merger timescales, where, at the highest redshifts ($z \gtrsim 3$), merging via direct collisional processes might be more efficient than merging via dynamical friction, given the large physical densities. More detailed study in very high-resolution numerical simulations will be necessary to determine the effective breakdown between different merger processes. Nevertheless, the difference in our predictions at these redshifts is still within the range of observational uncertainty. Ultimately, we find that our predictions are robust above masses $M_{gal} \gtrsim 10^{10} \, M_{\odot}$, regardless of these possible changes to our model, as the theoretical subhalo mass functions and empirical halo occupation models are reasonably well constrained in this regime.

In addition to these specific observational predictions and tests, our model allows us to examine the physical origins of the distribution of major mergers of different galaxy masses and types. For example, there is a naturally defined major-merger scale (host halo mass M_{halo}) for galaxies of mass M_{gal}—the "small group scale," only slightly larger than the average halo hosting a galaxy of mass M_{gal}. This is the scale at which the probability to accrete a second galaxy of comparable mass $\sim M_{gal}$ (fuel for a major merger) first becomes significant. At smaller (relative) halo masses, the probability that the halo hosts a galaxy as large as M_{gal} declines rapidly. At larger masses, the probability that the halo will merge with or accrete another halo hosting a comparable $\sim M_{gal}$ galaxy increases, but the efficiency of the merger of these galaxies declines rapidly. We stress that this small group scale is indeed small—the average small group halo will still host only 1 galaxy of mass $\sim M_{gal}$, and groups will only consist of 2–3 members of similar mass. We also note that this does not mean that mergers occur (in a global sense) on a specific scale, since the small group scale is dif-

ferent for different galaxy masses. In fact, a consequence of this model is that mergers occur in halos of all masses and in all environments (including field and even void environments), as is observed (Sol Alonso et al. 2006; Goto 2005; Hogg et al. 2006), although the characteristic masses and star formation histories of galaxies merging may reflect their different environments/halo masses. Similarly, our model allows us to accurately predict and understand the (relatively weak) evolution of the merger fraction with redshift, and the relative evolution in merger rates as a function of mass (evolution of the major merger mass functions). The clustering properties and dependence of merger rates on both large-scale and small-scale environment are natural consequences of the fundamentally local nature of mergers, and we study in detail the effects of environment on merger rates as a function of scale.

Having characterized mergers in this way, we examine the role that mergers play in triggering quasars. Even if there are other quasar "triggers" dominant at some luminosities/redshifts, it is difficult to imagine a scenario in which the strong nuclear gas inflows from a merger do not cause rapid, near Eddington-limited accretion and ultimately yield some kind of quasar—and indeed such activity is ubiquitous in late-stage mergers (Komossa et al. 2003; Alexander et al. 2005a; Borys et al. 2005; Brand et al. 2006). We therefore make the simple *Ansatz* that gas-rich, major mergers will produce quasars (but do, in principle, allow for other fueling mechanisms as well). This model, with just the contribution of mergers to the quasar luminosity density, is able to account for the observed quasar luminosity density from $z = 0-6$. The rise and fall of the luminosity density with redshift, as well as the shape and evolution of the quasar luminosity function, are accurately reproduced. This also yields predictions of the local black hole mass function, cosmic X-ray background (see Hopkins et al. 2006a), AGN fractions as a function of galaxy mass/luminosity and redshift, large-scale quasar clustering as a function of luminosity and redshift, small-scale quasar clustering excesses, quasar host galaxy colors, and infrared luminosity functions, all in good agreement with those observed. In particular, matching the history of the bolometric luminosity density of quasars requires no knowledge or assumptions about quasar duty cycles, light curves, or lifetimes, only our determination of the global mass density in gas-rich major mergers.

In our model, the sharp rise and fall of the quasar luminosity density over cosmic time is the product of several factors. At high redshifts, the buildup of BH mass from $z \gtrsim 6$ to $z \sim 2$ is due in part to the growth of galaxy and halo mass, as most galaxies are rapidly forming, and the galaxy mass density involved in major mergers steadily increases with time. The rise is steeper than that in, e.g., the global star formation rate density of the universe, as it tracks just the major merger history (effectively, at these redshifts, the rise in the density of relatively massive "small group" sized halos), as opposed to the global buildup of the (relatively lower mass) halos hosting the most rapidly star-forming galaxies. Below redshift $z \sim 2$, merger rates begin to decline for all galaxies, and the exhaustion of gas in evolved systems slows the growth of quasars in two ways. First, major mergers of relatively gas-poor disks create shallower central potential wells for the remnant spheroid (i.e., lower σ values), and as a consequence BH growth self-regulates at lower masses (Hopkins et al. 2007c), in agreement with the observed evolution of the BH-host correlations with redshift (e.g., Peng et al. 2006). Second, an increasing fraction of galaxies (especially around $\sim L_*$, where most of the mass density resides) have already undergone major mergers and exist as "quenched" spheroids (with very little remaining cold, rotationally supported gas) whose major mergers will not excite quasar activity. Recent high-resolution cosmological simulations that

attempt to resolve the relevant merger and feedback effects regulating BH growth (Sijacki et al. 2007; Di Matteo et al. 2007) further support this scenario, with the combination of these effects and, primarily, the merger history of the universe regulating BH growth (at least at redshifts $z \lesssim 6$). The product of these effects yields the observed steep rise and fall of the quasar population with respect to its peak at $z \sim 2$, in good agreement with the observations and in contrast with the substantially more extended global star formation history of the universe.

We compare this model to one in which quasar fueling is primarily driven by secular processes—i.e., disk instabilities, bars, harassment, or any process that operates in nonmerging, gas-rich systems. We demonstrate that there are a number of robust, qualitatively distinct predictions from these models, including the following:

Quasar clustering.—A merger-driven model accurately predicts the observed large-scale clustering of quasars (both at $\sim L_*$ and as a detailed function of luminosity) as a function of redshift for the observed range $z \sim 0.5-4$. The clustering is, at all these redshifts, precisely that predicted for "small group" halos in which major mergers of gas-rich galaxies should proceed most efficiently. It is well established empirically that quasar clustering traces a characteristic host halo mass (Porciani et al. 2004; Wake et al. 2004; Croom et al. 2005; Porciani & Norberg 2006; Myers et al. 2007a; da Angela et al. 2007; Coil et al. 2007; Shen et al. 2007; Hopkins et al. 2007d), and investigations of the quasar proximity effect reach a similar conclusion (Faucher-Giguere et al. 2008; Kim & Croft 2007; Nascimento Guimaraes et al. 2007). Comparing this to independent, direct measurements of the small group scale of $\sim L_*$ gas-rich galaxies, and to the small group scale inferred from a wide variety of different halo occupation models, we show in all cases that these trace the same mass. In contrast, the clustering of typical star-forming galaxies is somewhat weaker (as expected relative to their small group scale) and yields an underestimate of quasar clustering at moderate and high redshifts. Only at low redshifts ($z \lesssim 0.5$) is there reasonable consistency between the clustering of $\sim L_*$ quasars and "secular" populations (for more details, see Hopkins et al. 2007d).

Small-scale environments.—Mergers will preferentially occur in environments with an overdensity of galaxies on small scales, and as a consequence their clustering should reflect a bias (relative to a mean galaxy of the same mass) to excess clustering on small scales. Furthermore, triggering of binary quasars in (even a small fraction of) early interacting pairs can enhance this excess. Indeed, in a purely empirical sense, both bright quasars at all redshifts $z \sim 0.5-3$ (Hennawi et al. 2006; Serber et al. 2006; Myers et al. 2007b) and local poststarburst merger remnant galaxies (Goto 2005) are observed to have similar, strong excess clustering on small scales, distinct from quiescent (nonmerger related) populations. This is true both in terms of the quasar-quasar autocorrelation, and for the quasar-galaxy cross-correlation, suggesting that it reflects a true tendency for quasars to reside in regions of small-scale overdensity. Our model predicts the magnitude of this excess clustering as a function of physical scale and redshift well for both populations. Interestingly, low-luminosity Seyfert galaxies ($M_B > -23$) are observed without such an excess on small scales (Serber et al. 2006), as expected if AGN triggering at low luminosities (or typical $M_{\rm BH} \lesssim 10^7 \ M_\odot$) is dominated by secular processes (with the true quasar populations dominated by mergers). However, systems of these low luminosities contribute significantly to the quasar luminosity density at only very low redshifts $z \lesssim 0.5$, once more massive systems have predominantly quenched.

Host galaxy color.—The stellar population colors of a gas-rich merger remnant will rapidly redden, at least over the $\sim$ Gyr period over which subsequent infall or cooling can be ignored, and the system will (even if only temporarily) cross the "green valley" between the blue cloud and red sequence. If a quasar is triggered at the end of a merger, the decay of the quasar light curve should be associated with the host crossing this interval, or equivalently with the presence of a relatively young, blue host spheroid. Observed quasar hosts at $z \sim 0.5-1.1$ appear to preferentially occupy this (otherwise relatively empty) locus in color-magnitude space (Sánchez et al. 2004; Nandra et al. 2007), and it is well established that bright quasar hosts tend to be massive spheroids with especially young stellar or poststarburst stellar populations (e.g., Canalizo & Stockton 2001; Jahnke et al. 2004a; Vanden Berk et al. 2006; Barthel 2006 and references therein). We show that the color distribution of observed quasar hosts is similar to that observed for clear poststarburst merger remnant populations. In contrast, a secular model (regardless of the quasar duty cycle or lifetime) would predict that quasar hosts trace the population of systems hosting strong disk instabilities or bars (unless any quasar activity could somehow be suppressed over the entire lifetime of a relatively long-lived bar)—these actually tend to be the most blue, gas-rich disk galaxies. We show that the observed colors of quasar hosts are distinct from those of systems observed hosting strong bars.

Host kinematics (pseudobulges vs. classical bulges).—Numerical simulations and observations of both barred systems and merger remnants have established that mergers yield systems with the observed kinematic and photometric properties of classical bulges, whereas secular disk instabilities generically give rise to pseudobulges with distinct properties (see the discussion in § 1). At high redshifts $z \gtrsim 1$, the active $\sim L_*$ quasar populations (either from direct quasar BH mass measurements or simply the Eddington argument) are dominated by massive BHs ($M_{\rm BH} \gtrsim 10^8 \ M_\odot$), which are directly observed to live in massive bulges at those redshifts (Peng et al. 2006) and whose remnants clearly live in massive bulges locally. These spheroids ($M_{\rm sph} \gtrsim 10^{11} \ M_\odot$) are overwhelmingly classical spheroids (in particular, classical true ellipticals), whose kinematics argue that they were formed in mergers. To the extent that the buildup of BH mass traces spheroid origin (true at all redshifts observed, albeit with potentially redshift-dependent efficiency), this implies formation in mergers. Adopting a number of different estimates of, e.g., the pseudobulge fraction as a function of host properties, pseudobulge mass distributions, or simply assuming all bulges in star-forming/disk-dominated galaxies are formed via secular instabilities, we compare with the distribution of active BH masses in the quasar luminosity function at all redshifts and show that these populations cannot dominate the QLF at redshifts $z \gtrsim 1$. Only at low redshifts $z \lesssim 1$ are the global QLF and buildup of BH mass occurring mainly in systems that typically reside in star-forming, disk-dominated hosts with pseudobulges potentially formed via disk instabilities or bars.

Quasar luminosity density versus redshift.—As noted above, a merger-driven model predicts a sharp rise and fall of the quasar luminosity density in good agreement with observations. If, for the sake of argument, we adopt a model in which all BH growth is driven by disk instabilities, we demonstrate that, once embedded in a proper cosmological context, such a model is generically forced to predict a history of quasar luminosity density that is offset to earlier times (in each of its rise, peak, and fall), in conflict with the observations. This is because major mergers are dynamically inevitable—one cannot simply "remove" the mergers a galaxy will undergo in a true cosmological model. In order for disk instabilities to dominate BH growth or spheroid formation, they must,

therefore, act before massive systems undergo their major mergers. Since the global mass flux in gas-rich major mergers peaks around $z \sim 2$–3, a secular-dominant model is forced to assume a sufficiently strong disk instability mode such that the progenitors of these systems rapidly exhaust their gas supplies and build up most of their final BH/spheroid mass at redshifts $z \gtrsim 4$. By $z \sim 2$, then, these models predict the quasar luminosity density is already in rapid decline. We demonstrate this both for current state-of-the-art semianalytic models (Bower et al. 2006), constrained such that they cannot overproduce the $z = 0$ mass density in quenched systems nor "avoid" major mergers, and simple illustrative toy models. The only way to avoid this is to weaken the disk instability criterion—i.e., to assume disk instabilities are not so efficient at exhausting systems and can therefore act continuously over longer times. But then, one obtains a prediction similar to our expectation from assuming all pseudobulges are formed in disk instabilities—namely, the high rate of gas-rich mergers at high redshifts will dominate quasar activity at all $z > 1$, and this "gentler" disk instability mode will dominate at lower luminosities (i.e., only dominate BH mass buildup at low masses $M_{BH} \lesssim 10^7 \, M_\odot$), becoming important to the total luminosity density only at $z < 1$.

These comparisons, despite the very different possible systematic effects in the observations, all suggest a similar scenario. Secular (nonmerger related) fueling mechanisms may dominate AGN activity in low-BH mass systems ($M_{BH} \lesssim 10^7 \, M_\odot$), for which mergers are relatively rare and hosts tend to be very gas-rich, potentially bar-unstable disks, but these contribute little to quasar activity at $z \gtrsim 1$, which involves the most massive $M_{BH} \gtrsim 10^8 \, M_\odot$ BHs in the most massive spheroids. By $z \sim 0.5$, however, the most massive BHs are no longer active (their hosts having primarily been gas exhausted and quenched, and with overall merger rates declining), and a significant fraction of the AGN luminosity density can come from $\sim 10^7 \, M_\odot$ BHs in undisturbed hosts, corresponding to relatively low-luminosity ($M_B > -23$) Seyfert galaxies. By $z \sim 0$, the local QLF is largely dominated by Seyfert activity in relatively small BHs with late-type, undisturbed host disks (Heckman et al. 2004). Our models allow for secular mechanisms, such as the stochastic triggering model of Hopkins & Hernquist (2006) to be important at low luminosities, and a pure comparison between this secular model and our merger-driven prediction here yields a transition to secular dominance at low luminosities in good agreement with the empirical constraints.

Ultimately, one would like to test this by directly studying the morphology of true, bright quasar hosts at high redshifts. Unfortunately, as discussed in § 1, this remains extremely difficult, and results have been ambiguous. As noted previously, mock observations constructed from numerical major merger simulations (E. Krause et al. 2008, in preparation) imply that, with the best presently attainable data, the faint, rapidly fading tidal features associated with the quasar phase (i.e., final stages of the merger, at which the spheroid is largely formed and has begun to relax) are difficult to observe even locally and (for now) nearly impossible to identify at the redshifts of greatest interest ($z \gtrsim 1$). Similarly, experiments with automated, nonparametric classification

schemes (Lotz et al. 2004) suggest that the hosts will generically be classified as "normal" spheroids, even with perfect resolution and no surface brightness dimming. This appears to be borne out, as recently Bennert et al. (2008) have re-examined very low-redshift quasars previously recognized from deep *HST* imaging as having relaxed spheroid hosts and found (after considerably deeper integrations) that every such object shows clear evidence for a recent merger. The ability to identify such features may be slightly improved if one considers just the population of highly dust-reddened (but still dominated by quasar light in the optical/near IR) or IR-luminous quasars expected to be associated with a (brief) "blowout" stage preceding the more typical optical quasar phase in a merger, and it does appear that observations of quasars in this stage, somewhat closer to the peak of merger activity, show ubiquitous evidence of recent or ongoing mergers (Hutchings et al. 2003, 2006; Kawakatu et al. 2006; Guyon et al. 2006; Urrutia et al. 2008), albeit still requiring very deep integrations.

On the other hand, it is increasingly possible to improve the constraints we have studied in this paper, to break the degeneracy between secular and merger-driven models of quasar fueling. Improving measurements of merger fractions, mass functions, and clustering at low redshifts, and extending these measurements to high redshifts, can break the degeneracies in our cosmological models (regarding, for example, the appropriate merger timescales at high redshifts) and enable more robust, tightly constrained predictions. We have also made a large number of predictions in this paper and previous related works (e.g., Hopkins et al. 2006a, 2007d) that can be directly tested without the large ambiguities presently inherent in quasar host morphology estimates. Better observations of quasar host galaxy colors (and corresponding estimates of their recent star formation history), improved measurements of quasar clustering at redshifts $z \gtrsim 3$ (especially measurements that can resolve $\sim L_*$ quasars at these redshifts), detailed cross-correlation measurements of quasars and other galaxy populations and clustering measurements that can decompose the excess bias of quasars on small scales as a function of, e.g., redshift and luminosity, improved constraints on the bolometric corrections of the brightest quasars and the history of the bolometric quasar luminosity density at $z \gtrsim 3$–4, and estimates of the evolution with redshift of pseudobulge populations will all be able to test the models presented in this paper. The combination of these observations can greatly strengthen the constraints herein and ultimately allow for more detailed modeling that attempts not just to predict the general origin of quasars in mergers, but to fully break down the contribution of major mergers (or mergers of different types) and other fueling mechanisms to the quasar luminosity functions as a function of luminosity and redshift.

REFERENCES

Adelberger, K. L., & Steidel, C. C. 2005, ApJ, 630, 50
Adelberger, K. L., Steidel, C. C., Pettini, M., Shapley, A. E., Reddy, N. A., & Erb, D. K. 2005, ApJ, 619, 697
Alexander, D. M., Bauer, F. E., Chapman, S. C., Smail, I., Blain, A. W., Brandt, W. N., & Ivison, R. J. 2005a, ApJ, 632, 736
Alexander, D. M., Smail, I., Bauer, F. E., Chapman, S. C., Blain, A. W., Brandt, W. N., & Ivison, R. J. 2005b, Nature, 434, 738
Allen, P. D., Driver, S. P., Graham, A. W., Cameron, E., Liske, J., & de Propris, R. 2006, MNRAS, 371, 2

Allen, P. D., Moustakas, L. A., Dalton, G., MacDonald, E., Blake, C., Clewley, L., Heymans, C., & Wegner, G. 2005, MNRAS, 360, 1244
Alonso, M. S., Lambas, D. G., Tissera, P., & Coldwell, G. 2007, MNRAS, 375, 1017
Athanassoula, E. 2005, MNRAS, 358, 1477
Athanassoula, E., Bienayme, O., Martinet, L., & Pfenniger, D. 1983, A&A, 127, 349
Babbedge, T. S. R., et al. 2006, MNRAS, 370, 1159
Bahcall, J. N., Kirhakos, S., Saxe, D. H., & Schneider, D. P. 1997, ApJ, 479, 642

Balcells, M., Graham, A. W., & Peletier, R. F. 2007, ApJ, 665, 1104

Baldry, I. K., Balogh, M. L., Bower, R. G., Glazebrook, K., Nichol, R. C., Bamford, S. P., & Budavari, T. 2006, MNRAS, 373, 469

Ball, N. M., Loveday, J., Brunner, R. J., Baldry, I. K., & Brinkmann, J. 2006, MNRAS, 373, 845

Barazza, F. D., Jogee, S., & Marinova, I. 2007, in IAU Symp. 235, Galaxy Evolution across the Hubble Time, ed. F. Combes & J. Palous (Cambridge: Cambridge Univ. Press), 76

Barnes, J. E. 1988, ApJ, 331, 699

———. 1992, ApJ, 393, 484

Barnes, J. E., & Hernquist, L. 1992, ARA&A, 30, 705

———. 1996, ApJ, 471, 115

Barnes, J. E., & Hernquist, L. E. 1991, ApJ, 370, L65

Barthel, P. D. 2006, A&A, 458, 107

Bell, E. F., & de Jong, R. S. 2001, ApJ, 550, 212

Bell, E. F., McIntosh, D. H., Katz, N., & Weinberg, M. D. 2003a, ApJ, 585, L117

———. 2003b, ApJS, 149, 289

Bell, E. F., Phleps, S., Somerville, R. S., Wolf, C., Borch, A., & Meisenheimer, K. 2006, ApJ, 652, 270

Bell, E. F., Wolf, C., Meisenheimer, K., Rix, H.-W., Borch, A., Dye, S., Kleinheinrich, M., Wisotzki, L., & McIntosh, D. H. 2004, ApJ, 608, 752

Bell, E. F., et al. 2005, ApJ, 625, 23

Bennert, N., Canalizo, G., Jungwiert, B., Stockton, A., Schweizer, F., Peng, C. Y., & Lacy, M. 2008, ApJ, in press (arXiv: 0801.0832)

Binney, J., & Tremaine, S. 1987, Galactic Dynamics (Princeton: Princeton Univ. Press)

Blain, A. W., Chapman, S. C., Smail, I., & Ivison, R. 2004, ApJ, 611, 725

Blake, C., et al. 2004, MNRAS, 355, 713

Blanton, M. R. 2006, ApJ, 648, 268

Borch, A., et al. 2006, A&A, 453, 869

Borys, C., Smail, I., Chapman, S. C., Blain, A. W., Alexander, D. M., & Ivison, R. J. 2005, ApJ, 635, 853

Bournaud, F., Jog, C. J., & Combes, F. 2005, A&A, 437, 69

Bower, R. G., Benson, A. J., Malbon, R., Helly, J. C., Frenk, C. S., Baugh, C. M., Cole, S., & Lacey, C. G. 2006, MNRAS, 370, 645

Brand, K., et al. 2006, ApJ, 644, 143

———. 2007, ApJ, 663, 204

Bridge, C. R., et al. 2007, ApJ, 659, 931

Brotherton, M. S., et al. 1999, ApJ, 520, L87

Brough, S., Forbes, D. A., Kilborn, V. A., & Couch, W. 2006, MNRAS, 370, 1223

Bundy, K., Ellis, R. S., & Conselice, C. J. 2005, ApJ, 625, 621

Bundy, K., Fukugita, M., Ellis, R. S., Kodama, T., & Conselice, C. J. 2004, ApJ, 601, L123

Bundy, K., et al. 2006, ApJ, 651, 120

Canalizo, G., & Stockton, A. 2001, ApJ, 555, 719

Caputi, K. I., et al. 2007, ApJ, 660, 97

Carollo, C. M., Stiavelli, M., & Mack, J. 1998, AJ, 116, 68

Cassata, P., et al. 2005, MNRAS, 357, 903

Chakrabarti, S., Cox, T. J., Hernquist, L., Hopkins, P. F., Robertson, B., & Di Matteo, T. 2007, ApJ, 658, 840

Chapman, S. C., Blain, A. W., Smail, I., & Ivison, R. J. 2005, ApJ, 622, 772

Chary, R., & Elbaz, D. 2001, ApJ, 556, 562

Coil, A. L., Hennawi, J. F., Newman, J. A., Cooper, M. C., & Davis, M. 2007, ApJ, 654, 115

Coil, A. L., et al. 2004, ApJ, 609, 525

———. 2006, ApJ, 638, 668

Combes, F., Debbasch, F., Friedli, D., & Pfenniger, D. 1990, A&A, 233, 82

Conroy, C., Wechsler, R. H., & Kravtsov, A. V. 2006, ApJ, 647, 201

Conroy, C., et al. 2007, ApJ, 654, 153

Conselice, C. J., Bershady, M. A., Dickinson, M., & Papovich, C. 2003, AJ, 126, 1183

Cooray, A. 2005, MNRAS, 364, 303

———. 2006, MNRAS, 365, 842

Cox, T. J., Di Matteo, T., Hernquist, L., Hopkins, P. F., Robertson, B., & Springel, V. 2006a, ApJ, 643, 692

Cox, T. J., Dutta, S. N., Di Matteo, T., Hernquist, L., Hopkins, P. F., Robertson, B., & Springel, V. 2006b, ApJ, 650, 791

Croom, S. M., Boyle, B. J., Shanks, T., Smith, R. J., Miller, L., Outram, P. J., Loaring, N. S., Hoyle, F., & da Ângela, J. 2005, MNRAS, 356, 415

Croton, D. J., Gao, L., & White, S. D. M. 2007, MNRAS, 374, 1303

Croton, D. J., et al. 2006, MNRAS, 365, 11

da Angela, J., et al. 2008, MNRAS, 383, 565

Dasyra, K. M., et al. 2006, ApJ, 638, 745

———. 2007, ApJ, 657, 102

De Lucia, G., Kauffmann, G., Springel, V., White, S. D. M., Lanzoni, B., Stoehr, F., Tormen, G., & Yoshida, N. 2004, MNRAS, 348, 333

De Propris, R., Liske, J., Driver, S. P., Allen, P. D., & Cross, N. J. G. 2005, AJ, 130, 1516

Di Matteo, T., Colberg, J., Springel, V., Hernquist, L., & Sijacki, D. 2007, ApJ, submitted (arXiv:0705.2269v1)

Di Matteo, T., Springel, V., & Hernquist, L. 2005, Nature, 433, 604

Djorgovski, S., & Davis, M. 1987, ApJ, 313, 59

Dressler, A., Lynden-Bell, D., Burstein, D., Davies, R. L., Faber, S. M., Terlevich, R., & Wegner, G. 1987, ApJ, 313, 42

Driver, S. P., Allen, P. D., Liske, J., & Graham, A. W. 2007, ApJ, 657, L85

Eke, V. R., et al. 2004, MNRAS, 355, 769

Elvis, M., et al. 1994, ApJS, 95, 1

Erb, D. K., Steidel, C. C., Shapley, A. E., Pettini, M., Reddy, N. A., & Adelberger, K. L. 2006, ApJ, 646, 107

Falomo, R., Kotilainen, J. K., Pagani, C., Scarpa, R., & Treves, A. 2004, ApJ, 604, 495

Fan, X., et al. 2004, AJ, 128, 515

Farrah, D., Afonso, J., Efstathiou, A., Rowan-Robinson, M., Fox, M., & Clements, D. 2003, MNRAS, 343, 585

Faucher-Giguere, C., Lidz, A., Zaldarriaga, M., & Hernquist, L. 2008, 673, 39

Ferrarese, L., & Merritt, D. 2000, ApJ, 539, L9

Fine, S., et al. 2006, MNRAS, 373, 613

Floyd, D. J. E., Kukula, M. J., Dunlop, J. S., McLure, R. J., Miller, L., Percival, W. J., Baum, S. A., & O'Dea, C. P. 2004, MNRAS, 355, 196

Fontana, A., et al. 2004, A&A, 424, 23

———. 2006, A&A, 459, 745

Franceschini, A., et al. 2005, AJ, 129, 2074

———. 2006, A&A, 453, 397

Furlanetto, S. R., & Kamionkowski, M. 2006, MNRAS, 366, 529

Gao, L., & White, S. D. M. 2007, MNRAS, 377, L5

Gao, L., White, S. D. M., Jenkins, A., Stoehr, F., & Springel, V. 2004, MNRAS, 355, 819

Gebhardt, K., et al. 2000, ApJ, 539, L13

Giavalisco, M., & Dickinson, M. 2001, ApJ, 550, 177

Gilli, R., Comastri, A., & Hasinger, G. 2007, A&A, 463, 79

Goto, T. 2005, MNRAS, 357, 937

———. 2006, MNRAS, 369, 1765

Gottlöber, S., Klypin, A., & Kravtsov, A. V. 2001, ApJ, 546, 223

Graham, A. W., & Driver, S. P. 2007, ApJ, 655, 77

Graham, A. W., Erwin, P., Caon, N., & Trujillo, I. 2001, ApJ, 563, L11

Granato, G. L., De Zotti, G., Silva, L., Bressan, A., & Danese, L. 2004, ApJ, 600, 580

Grazian, A., Negrello, M., Moscardini, L., Cristiani, S., Haehnelt, M. G., Matarrese, S., Omizzolo, A., & Vanzella, E. 2004, AJ, 127, 592

Gregg, M. D., Lacy, M., White, R. L., Glikman, E., Helfand, D., Becker, R. H., & Brotherton, M. S. 2002, ApJ, 564, 133

Greene, J. E., & Ho, L. C. 2007, ApJ, 667, 131

Guyon, O., Sanders, D. B., & Stockton, A. 2006, ApJS, 166, 89

Hao, L., et al. 2005, AJ, 129, 1795

Häring, N., & Rix, H.-W. 2004, ApJ, 604, L89

Harker, G., Cole, S., Helly, J., Frenk, C., & Jenkins, A. 2006, MNRAS, 367, 1039

Hasinger, G., Miyaji, T., & Schmidt, M. 2005, A&A, 441, 417

Heckman, T. M., Kauffmann, G., Brinchmann, J., Charlot, S., Tremonti, C., & White, S. D. M. 2004, ApJ, 613, 109

Hennawi, J. F., et al. 2006, AJ, 131, 1

Hernquist, L. 1989, Nature, 340, 687

———. 1992, ApJ, 400, 460

———. 1993, ApJ, 409, 548

Hernquist, L., & Barnes, J. E. 1991, Nature, 354, 210

Hernquist, L., & Mihos, J. C. 1995, ApJ, 448, 41

Hernquist, L., & Quinn, P. J. 1987, ApJ, 312, 1

Hernquist, L., & Spergel, D. N. 1992, ApJ, 399, L117

Heymans, C., et al. 2006, MNRAS, 371, L60

Hogg, D. W., Masjedi, M., Berlind, A. A., Blanton, M. R., Quintero, A. D., & Brinkmann, J. 2006, ApJ, 650, 763

Hopkins, A. M. 2004, ApJ, 615, 209

Hopkins, A. M., & Beacom, J. F. 2006, ApJ, 651, 142

Hopkins, P. F., Bundy, K., Hernquist, L., & Ellis, R. S. 2007a, ApJ, 659, 976

Hopkins, P. F., Cox, T. J., Kereš, D., & Hernquist, L. 2007b, ApJ, submitted (Paper II)

Hopkins, P. F., & Hernquist, L. 2006, ApJS, 166, 1

Hopkins, P. F., Hernquist, L., Cox, T. J., Di Matteo, T., Martini, P., Robertson, B., & Springel, V. 2005a, ApJ, 630, 705

Hopkins, P. F., Hernquist, L., Cox, T. J., Di Matteo, T., Robertson, B., & Springel, V. 2005b, ApJ, 630, 716

———. 2005c, ApJ, 632, 81

———. 2006a, ApJS, 163, 1

Hopkins, P. F., Hernquist, L., Cox, T. J., Robertson, B., Di Matteo, T., & Springel, V. 2006b, ApJ, 639, 700

Hopkins, P. F., Hernquist, L., Cox, T. J., Robertson, B., & Krause, E. 2007c, ApJ, 669, 45

Hopkins, P. F., Hernquist, L., Cox, T. J., Robertson, B., & Springel, V. 2006c, ApJS, 163, 50

Hopkins, P. F., Hernquist, L., Martini, P., Cox, T. J., Robertson, B., Di Matteo, T., & Springel, V. 2005d, ApJ, 625, L71

Hopkins, P. F., Lidz, A., Hernquist, L., Coil, A. L., Myers, A. D., Cox, T. J., & Spergel, D. N. 2007d, ApJ, 662, 110

Hopkins, P. F., Narayan, R., & Hernquist, L. 2006d, ApJ, 643, 641

Hopkins, P. F., Richards, G. T., & Hernquist, L. 2007e, ApJ, 654, 731

Hopkins, P. F., Robertson, B., Krause, E., Hernquist, L., & Cox, T. J. 2006e, ApJ, 652, 107

Hopkins, P. F., Somerville, R. S., Hernquist, L., Cox, T. J., Robertson, B., & Li, Y. 2006f, ApJ, 652, 864

Hopkins, P. F., et al. 2004, AJ, 128, 1112

Hutchings, J. B. 2003, AJ, 125, 1053

Hutchings, J. B., Cherniawsky, A., Cutri, R. M., & Nelson, B. O. 2006, AJ, 131, 680

Hutchings, J. B., Maddox, N., Cutri, R. M., & Nelson, B. O. 2003, AJ, 126, 63

Infante, L., et al. 2002, ApJ, 567, 155

Jahnke, K., Kuhlbrodt, B., & Wisotzki, L. 2004a, MNRAS, 352, 399

Jahnke, K., et al. 2004b, ApJ, 614, 568

Jesseit, R., Naab, T., Peletier, R., & Burkert, A. 2007, MNRAS, 376, 997

Jogee, S. 2006, in Physics of Active Galactic Nuclei on All Scales, ed. D. Alloin, R. Johnson, & P. Lira (Berlin: Springer), 143

Kauffmann, G., & Haehnelt, M. 2000, MNRAS, 311, 576

———. 2002, MNRAS, 332, 529

Kauffmann, G., et al. 2003, MNRAS, 346, 1055

———. 2004, MNRAS, 353, 713

Kawakatu, N., Anabuki, N., Nagao, T., Umemura, M., & Nakagawa, T. 2006, ApJ, 637, 104

Kennicutt, R. C., Jr. 1998, ApJ, 498, 541

Kim, Y.-R., & Croft, R. 2007, MNRAS, submitted (astro-ph/0701012)

Kollmeier, J. A., et al. 2006, ApJ, 648, 128

Komossa, S., Burwitz, V., Hasinger, G., Predehl, P., Kaastra, J. S., & Ikebe, Y. 2003, ApJ, 582, L15

Kormendy, J., & Gebhardt, K. 2001, in AIP Conf. Proc. 586, 20th Texas Symposium on Relativistic Astrophysics, ed. J. C. Wheeler & H. Martel (New York: AIP), 363

Kormendy, J., & Kennicutt, R. C., Jr. 2004, ARA&A, 42, 603

Kormendy, J., & Richstone, D. 1995, ARA&A, 33, 581

Kravtsov, A. V., Berlind, A. A., Wechsler, R. H., Klypin, A. A., Gottlöber, S., Allgood, B., & Primack, J. R. 2004, ApJ, 609, 35

Kriek, M., et al. 2007, ApJ, 669, 776

Krivitsky, D. S., & Kontorovich, V. M. 1997, A&A, 327, 921

Kuijken, K., & Merrifield, M. R. 1995, ApJ, 443, L13

Lapi, A., Shankar, F., Mao, J., Granato, G. L., Silva, L., De Zotti, G., & Danese, L. 2006, ApJ, 650, 42

Le Floc'h, E., et al. 2005, ApJ, 632, 169

Lee, K.-S., Giavalisco, M., Gnedin, O. Y., Somerville, R. S., Ferguson, H. C., Dickinson, M., & Ouchi, M. 2006, ApJ, 642, 63

Letawe, G., Magain, P., Courbin, F., Jablonka, P., Jahnke, K., Meylan, G., & Wisotzki, L. 2007, MNRAS, 378, 83

Li, C., Kauffmann, G., Jing, Y. P., White, S. D. M., Börner, G., & Cheng, F. Z. 2006, MNRAS, 368, 21

Li, Y., et al. 2007, ApJ, 665, 187

———. 2008, ApJ, in press (arXiv: 0706.3706)

Lidz, A., Hopkins, P. F., Cox, T. J., Hernquist, L., & Robertson, B. 2006, ApJ, 641, 41

Lidz, A., McQuinn, M., Zaldarriaga, M., Hernquist, L., & Dutta, S. 2007, ApJ, 670, 39

Lin, L., et al. 2004, ApJ, 617, L9

Lotz, J. M., Madau, P., Giavalisco, M., Primack, J., & Ferguson, H. C. 2006, ApJ, 636, 592

Lotz, J. M., Primack, J., & Madau, P. 2004, AJ, 128, 163

Lotz, J. M., et al. 2008, ApJ, 672, 177

Lutz, D., Spoon, H. W. W., Rigopoulou, D., Moorwood, A. F. M., & Genzel, R. 1998, ApJ, 505, L103

Lynden-Bell, D. 1967, MNRAS, 136, 101

———. 1969, Nature, 223, 690

Magorrian, J., et al. 1998, AJ, 115, 2285

Makino, J., & Hut, P. 1997, ApJ, 481, 83

Maller, A. H., Katz, N., Kereš, D., Davé, R., & Weinberg, D. H. 2006, ApJ, 647, 763

Mamon, G. A. 2006, in Groups of Galaxies in the Nearby Universe, ed. I. Saviane, V. Ivanov, & J. Borissova (Garching: ESO), 203

Marconi, A., & Hunt, L. K. 2003, ApJ, 589, L21

Marconi, A., Risaliti, G., Gilli, R., Hunt, L. K., Maiolino, R., & Salvati, M. 2004, MNRAS, 351, 169

Martin, D. C., et al. 2007, ApJS, 173, 342

Martínez-Sansigre, A., Rawlings, S., Lacy, M., Fadda, D., Jarvis, M. J., Marleau, F. R., Simpson, C., & Willott, C. J. 2006, MNRAS, 370, 1479

Martini, P. 2004, in Coevolution of Black Holes and Galaxies, ed. L. C. Ho (Cambridge: Cambridge Univ. Press), 169

Masjedi, M., et al. 2006, ApJ, 644, 54

McLure, R. J., & Dunlop, J. S. 2002, MNRAS, 331, 795

———. 2004, MNRAS, 352, 1390

Menanteau, F., Ford, H. C., Motta, V., Benítez, N., Martel, A. R., Blakeslee, J. P., & Infante, L. 2006, AJ, 131, 208

Meneux, B., et al. 2006, A&A, 452, 387

Mihos, J. C., & Hernquist, L. 1994a, ApJ, 437, L47

———. 1994b, ApJ, 431, L9

———. 1996, ApJ, 464, 641

Mo, H. J., & White, S. D. M. 1996, MNRAS, 282, 347

Myers, A. D., Brunner, R. J., Nichol, R. C., Richards, G. T., Schneider, D. P., & Bahcall, N. A. 2007a, ApJ, 658, 85

Myers, A. D., Brunner, R. J., Richards, G. T., Nichol, R. C., Schneider, D. P., & Bahcall, N. A. 2007b, ApJ, 658, 99

Myers, A. D., et al. 2006, ApJ, 638, 622

Naab, T., & Burkert, A. 2003, ApJ, 597, 893

Naab, T., Burkert, A., & Hernquist, L. 1999, ApJ, 523, L133

Naab, T., Jesseit, R., & Burkert, A. 2006a, MNRAS, 372, 839

Naab, T., Khochfar, S., & Burkert, A. 2006b, ApJ, 636, L81

Naab, T., & Trujillo, I. 2006, MNRAS, 369, 625

Nandra, K., et al. 2007, ApJ, 660, L11

Nascimento Guimaraes, R., Petitjean, P., Beaumont Rollinde, E., Ramos De Carvalho, R., Djorgovski, G., Srianand, R., Aghaee, A., & Castro, S. 2007, MNRAS, 377, 657

Noeske, K. G., et al. 2007, ApJ, 660, L47

Noordermeer, E., & van der Hulst, J. M. 2007, MNRAS, 376, 1480

Norberg, P., et al. 2002, MNRAS, 332, 827

Nurmi, P., Heinämäki, P., Saar, E., Einasto, M., Holopainen, J., Martínez, V. J., & Einasto, J. 2008, A&A, in press (astro-ph/0611941)

O'Neill, J. K., & Dubinski, J. 2003, MNRAS, 346, 251

Patton, D. R., et al. 2002, ApJ, 565, 208

Peng, C. Y., Impey, C. D., Rix, H.-W., Kochanek, C. S., Keeton, C. R., Falco, E. E., Lehár, J., & McLeod, B. A. 2006, ApJ, 649, 616

Percival, W. J., Scott, D., Peacock, J. A., & Dunlop, J. S. 2003, MNRAS, 338, L31

Pérez-González, P. G., et al. 2005, ApJ, 630, 82

Pfenniger, D. 1984, A&A, 134, 373

Phleps, S., Peacock, J. A., Meisenheimer, K., & Wolf, C. 2006, A&A, 457, 145

Pierce, C. M., et al. 2007, ApJ, 660, L19

Polletta, M. d. C., et al. 2006, ApJ, 642, 673

Porciani, C., Magliocchetti, M., & Norberg, P. 2004, MNRAS, 355, 1010

Porciani, C., & Norberg, P. 2006, MNRAS, 371, 1824

Quinn, P. J. 1984, ApJ, 279, 596

Raha, N., Sellwood, J. A., James, R. A., & Kahn, F. D. 1991, Nature, 352, 411

Reed, D. S., Bower, R., Frenk, C. S., Jenkins, A., & Theuns, T. 2007, MNRAS, 374, 2

Richards, G. T., et al. 2003, AJ, 126, 1131

———. 2005, MNRAS, 360, 839

———. 2006a, ApJS, 166, 470

———. 2006b, AJ, 131, 2766

Richstone, D., et al. 1998, Nature, 395, A14

Robertson, B., Bullock, J. S., Cox, T. J., Di Matteo, T., Hernquist, L., Springel, V., & Yoshida, N. 2006a, ApJ, 645, 986

Robertson, B., Cox, T. J., Hernquist, L., Franx, M., Hopkins, P. F., Martini, P., & Springel, V. 2006b, ApJ, 641, 21

Robertson, B., Hernquist, L., Cox, T. J., Di Matteo, T., Hopkins, P. F., Martini, P., & Springel, V. 2006c, ApJ, 641, 90

Rodighiero, G., et al. 2007, MNRAS, 376, 416

Rothberg, B., & Joseph, R. D. 2006a, AJ, 131, 185

———. 2006b, AJ, 132, 976

Sajina, A., Yan, L., Armus, L., Choi, P., Fadda, D., Helou, G., & Spoon, H. 2007, ApJ, 664, 713

Salpeter, E. E. 1964, ApJ, 140, 796

Salucci, P., Szuszkiewicz, E., Monaco, P., & Danese, L. 1999, MNRAS, 307, 637

Salviander, S., Shields, G. A., Gebhardt, K., & Bonning, E. W. 2006, NewA Rev., 50, 803

Sánchez, S. F., et al. 2004, ApJ, 614, 586

Sanders, D. B., & Mirabel, I. F. 1996, ARA&A, 34, 749

Sanders, D. B., Soifer, B. T., Elias, J. H., Madore, B. F., Matthews, K., Neugebauer, G., & Scoville, N. Z. 1988a, ApJ, 325, 74

———. 1988b, ApJ, 325, 74

Sanders, D. B., Soifer, B. T., Elias, J. H., Neugebauer, G., & Matthews, K. 1988c, ApJ, 328, L35

Saunders, W., Rowan-Robinson, M., Lawrence, A., Efstathiou, G., Kaiser, N., Ellis, R. S., & Frenk, C. S. 1990, MNRAS, 242, 318
Schwarz, M. P. 1981, ApJ, 247, 77
Schweizer, F. 1992, in Physics of Nearby Galaxies: Nature or Nurture?, ed. T. X. Thuan, C. Balkowski, & J. Tran Thanh van (Gif-sur-Yvette: Editions Frontiéres), 283
Schweizer, F. 1996, AJ, 111, 109
———. 1998, in Galaxies: Interactions and Induced Star Formation, ed. R. C. Kennicutt Jr., F. Schweizer, J. E. Barnes, D. Friedli, L. Martinet, & D. Pfenniger (Berlin: Springer), 105
Schweizer, F., & Seitzer, P. 1992, AJ, 104, 1039
Serber, W., Bahcall, N., Ménard, B., & Richards, G. 2006, ApJ, 643, 68
Shakura, N. I., & Sunyaev, R. A. 1973, A&A, 24, 337
Shankar, F., Salucci, P., Granato, G. L., De Zotti, G., & Danese, L. 2004, MNRAS, 354, 1020
Shaw, L. D., Weller, J., Ostriker, J. P., & Bode, P. 2006, ApJ, 646, 815
Shen, S., Mo, H. J., White, S. D. M., Blanton, M. R., Kauffmann, G., Voges, W., Brinkmann, J., & Csabai, I. 2003, MNRAS, 343, 978
Shen, Y., et al. 2007, AJ, 133, 2222
Shepherd, C. W., et al. 2001, ApJ, 560, 72
Sheth, R. K., Mo, H. J., & Tormen, G. 2001, MNRAS, 323, 1
Shields, G. A., Gebhardt, K., Salviander, S., Wills, B. J., Xie, B., Brotherton, M. S., Yuan, J., & Dietrich, M. 2003, ApJ, 583, 124
Shields, G. A., Menezes, K. L., Massart, C. A., & Vanden Bout, P. 2006, ApJ, 641, 683
Sijacki, D., Springel, V., Di Matteo, T., & Hernquist, L. 2007, MNRAS, 380, 877
Simões Lopes, R. D., Storchi-Bergmann, T., de Fátima Saraiva, M., & Martini, P. 2007, ApJ, 655, 718
Smith, R. E., et al. 2003, MNRAS, 341, 1311
Soifer, B. T., & Neugebauer, G. 1991, AJ, 101, 354
Sol Alonso, M., Lambas, D. G., Tissera, P., & Coldwell, G. 2006, MNRAS, 367, 1029
Soltan, A. 1982, MNRAS, 200, 115
Spergel, D. N., et al. 2003, ApJS, 148, 175
———. 2007, ApJS, 170, 377
Springel, V., Di Matteo, T., & Hernquist, L. 2005a, ApJ, 620, L79
———. 2005b, MNRAS, 361, 776
Springel, V., & Hernquist, L. 2003, MNRAS, 339, 289
Springel, V., White, S. D. M., Tormen, G., & Kauffmann, G. 2001, MNRAS, 328, 726
Springel, V., et al. 2005c, Nature, 435, 629
Stevens, J. A., Page, M. J., Ivison, R. J., Carrera, F. J., Mittaz, J. P. D., Smail, I., & McHardy, I. M. 2005, MNRAS, 360, 610
Strateva, I., et al. 2001, AJ, 122, 1861
Straughn, A. N., Cohen, S. H., Ryan, R. E., Hathi, N. P., Windhorst, R. A., & Jansen, R. A. 2006, ApJ, 639, 724
Taylor, J. E., & Babul, A. 2004, MNRAS, 348, 811
Thacker, R. J., Scannapieco, E., & Couchman, H. M. P. 2006, ApJ, 653, 86
Tinker, J. L., Weinberg, D. H., Zheng, Z., & Zehavi, I. 2005, ApJ, 631, 41

Toomre, A. 1977, in Evolution of Galaxies and Stellar Populations, ed. B. M. Tinsley & R. B. Larson (New Haven: Yale Univ. Obs.), 401
Toomre, A., & Toomre, J. 1972, ApJ, 178, 623
Tormen, G., Moscardini, L., & Yoshida, N. 2004, MNRAS, 350, 1397
Tremaine, S., et al. 2002, ApJ, 574, 740
Ueda, Y., Akiyama, M., Ohta, K., & Miyaji, T. 2003, ApJ, 598, 886
Urrutia, T., Lacy, M., & Becker, R. H. 2008, ApJ, 674, 80
Vale, A., & Ostriker, J. P. 2006, MNRAS, 371, 1173
van den Bosch, F. C., Tormen, G., & Giocoli, C. 2005, MNRAS, 359, 1029
van den Bosch, F. C., et al. 2007, MNRAS, 376, 841
Vanden Berk, D. E., et al. 2006, AJ, 131, 84
Vestergaard, M. 2004, ApJ, 601, 676
Volonteri, M., Salvaterra, R., & Haardt, F. 2006, MNRAS, 373, 121
Wake, D. A., Miller, C. J., Di Matteo, T., Nichol, R. C., Pope, A., Szalay, A. S., Gray, A., Schneider, D. P., & York, D. G. 2004, ApJ, 610, L85
Walter, F., Carilli, C., Bertoldi, F., Menten, K., Cox, P., Lo, K. Y., Fan, X., & Strauss, M. A. 2004, ApJ, 615, L17
Wang, L., Li, C., Kauffmann, G., & de Lucia, G. 2006, MNRAS, 371, 537
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
White, R. L., Helfand, D. J., Becker, R. H., Gregg, M. D., Postman, M., Lauer, T. R., & Oegerle, W. 2003, AJ, 126, 706
White, S. D. M. 1976, MNRAS, 174, 467
Wolf, C., et al. 2005, ApJ, 630, 771
Woo, J.-H., Treu, T., Malkan, M. A., & Blandford, R. D. 2006, ApJ, 645, 900
Woods, D. F., Geller, M. J., & Barton, E. J. 2006, AJ, 132, 197
Wyithe, J. S. B., & Loeb, A. 2002, ApJ, 581, 886
Xu, C. K., Sun, Y. C., & He, X. T. 2004, ApJ, 603, L73
Yan, R., Madgwick, D. S., & White, M. 2003, ApJ, 598, 848
Yang, X., Mo, H. J., & van den Bosch, F. C. 2003, MNRAS, 339, 1057
Yang, Y., Tremonti, C. A., Zabludoff, A. I., & Zaritsky, D. 2006, ApJ, 646, L33
Yang, Y., Zabludoff, A. I., Zaritsky, D., Lauer, T. R., & Mihos, J. C. 2004, ApJ, 607, 258
Yip, C. W., et al. 2004, AJ, 128, 2603
Yu, Q., & Lu, Y. 2004, ApJ, 602, 603
Yu, Q., & Tremaine, S. 2002, MNRAS, 335, 965
Yun, M. S., Reddy, N. A., & Condon, J. J. 2001, ApJ, 554, 803
Zakamska, N. L., Strauss, M. A., Heckman, T. M., Ivezić, Ž., & Krolik, J. H. 2004, AJ, 128, 1002
Zakamska, N. L., et al. 2006, AJ, 132, 1496
Zehavi, I., et al. 2005, ApJ, 630, 1
Zentner, A. R., Berlind, A. A., Bullock, J. S., Kravtsov, A. V., & Wechsler, R. H. 2005, ApJ, 624, 505
Zheng, Z., et al. 2005, ApJ, 633, 791