The Early Evolution of Massive Black Holes

Marta Volonteri

1Astronomy Department, University of Michigan
500 Church Street, Ann Arbor, MI 48109, USA
Email: martav@umich.edu

Abstract. Massive black holes (MBHs) are nowadays believed to reside in most local galaxies. Studies have also established a number of relations between the MBH mass and properties of the host galaxy such as bulge mass and velocity dispersion. These results suggest that central MBHs, while much less massive than their hosts (~ 0.1%), are linked to the evolution of galactic structure. When did it all start? In hierarchical cosmologies, a single big galaxy today can be traced back to the stage when it was split up in hundreds of smaller components. Did MBH seeds form with the same efficiency in small proto-galaxies, or did their formation have to await the buildup of substantial galaxies with deeper potential wells? I briefly review here some of the physical processes that are conducive to the evolution of the massive black hole population. I will discuss black hole formation processes for “seed” black holes that are likely to take place at early cosmic epochs, and possible observational tests of these scenarios.

Keywords. cosmology: theory, galaxies: formation, gravitation, black hole physics, quasars: general

1. Introduction

I will focus here on the formation and evolution of massive black holes (MBHs) in high-redshift galaxies, and their symbiotic evolution with their hosts. MBHs weighing a million solar masses and above have been recognized as the engines that power quasars detected at early cosmic times. Dynamical evidence also indicates that MBHs with masses in the range \( M_{\text{BH}} \sim 10^6 - 10^9 M_\odot \) ordinarily dwell in the centers of most nearby galaxies (Ferrarese & Ford 2005). MBHs populate galaxy centers today, and shone as quasars in the past; the quiescent MBHs that we detect now in nearby bulges could be the dormant remnants of this fiery past, suggesting a single mechanism for assembling MBHs and forming galaxies. The surprisingly clear correlations between MBH masses and the properties of their host galaxies (Gültekin et al. 2009, and references therein) may well extend down to the smallest masses; for example, the dwarf Seyfert 1 galaxies POX 52 and NGC 4395 are each thought to contain a MBH of mass \( M_{\text{BH}} \sim 10^5 M_\odot \) (Barth et al. 2004; Peterson et al. 2005). At the other end, however, some powerful quasars, powered by supermassive black holes with masses \( \sim 10^9 M_\odot \) (Barth et al. 2003), have already been detected at \( z > 6 \), corresponding to a time less than a tenth of the age of the universe, roughly one billion years after the Big Bang.

We are therefore left with the task of explaining the presence of billion solar-mass MBHs when the universe is less than 1 Gyr old, and of much smaller MBHs lurking in 13 Gyr-old galaxies. The outstanding questions concern how and when “seed” MBHs formed, the frequency of MBHs in galaxies, and how efficiently MBH seeds grew in mass during the first few billion years of their lives.

The “flow chart” presented by Rees (1978) still stands as a guideline for the possible paths leading to formation of massive MBH seeds in the center of galactic structures. In this paper I will concentrate on the currently favored scenarios.

2.1. Population III Remnants

One of the most popular scenarios for MBH formation associates MBH seeds with the remnants of the first generation of stars, formed out of zero metallicity gas. The first stars are expected to form in minihalos, \( M_{\text{min}} \approx 10^6 M_\odot \) collapsing at \( z \approx 20 - 50 \) from the highest peaks of the primordial density field, where cooling is possible by means of molecular hydrogen (Palla \textit{et al.} 2002). Simulations of the collapse of primordial molecular clouds (Bromm \textit{et al.} 1999, 2002; Abel \textit{et al.} 2000; Yoshida \textit{et al.} 2006) suggest that the first generation of stars contained many “very massive stars” (VMSs) with \( M_* > 100 M_\odot \).

This is because of the slow subsonic contraction set up by molecular hydrogen cooling. Further fragmentation into sub-components is not seen, although it is not clear if this is a numerical effect rather than due to the gas physics (Glover \textit{et al.} 2008).

If the first stars retain their high mass until death, they will collapse after a short (of order Myrs) lifetime. Their final fate depends on the exact mass of the star. Between 25 and 140 \( M_\odot \), low-metallicity stars form black holes directly. It is likely, however, that such light black holes wander within their hosts, dynamically interacting with stars of similar mass, rather than settling at the center of the galaxy’s potential well. Between approximately 140 and 260 \( M_\odot \) lies the domain of pair-instability supernovae. These objects are completely disrupted by nuclear-powered explosions, leaving no remnants (Kudritzki & Puls 2000). Going to still more massive stars (over 260 \( M_\odot \)) the nuclear energy released when the star collapses on the pair instability is not sufficient to reverse the implosion. A massive black hole, containing at least half of the initial stellar mass, is born inside the star (Bond \textit{et al.} 1984; Fryer \textit{et al.} 2001). A numerous population of MBHs may have been the end–product of the first episode of pregalactic star formation; since they form in high-density peaks, relic MBHs with mass \( \gtrsim 150 M_\odot \) would be predicted to cluster in the cores of more massive halos formed by subsequent mergers (Madau & Rees 2001).

Although this path to MBH formation seems very natural, large uncertainties exist on the final mass of Population III stars. If the first stars formed in multiples per halo, the mass function would be less top-heavy (Glover \textit{et al.} 2008). And if single objects are formed, the characteristic final mass at the end of the accretion process could be much less than the initial clump mass due to feedback effects (McKee & Tan 2008).

2.2. Gas-Dynamical Processes

Another family of models for MBH formation relies on the collapse of supermassive objects formed directly out of dense gas (e.g., Bromm & Loeb 2003; Begelman \textit{et al.} 2006; Lodato & Natarajan 2006). The physical conditions (density, gas content) in the inner regions of mainly gaseous proto-galaxies make these loci natural candidates, because the very first proto-galaxies were by definition metal-free, or at the very least very metal-poor. Enriched halos have more efficient cooling, which in turn favors fragmentation and star formation over the efficient collection of gas conducive to MBH formation. It has been suggested that efficient gas collapse probably occurs only in massive halos with virial temperatures \( T_{\text{vir}} \gtrsim 10^4 \text{ K} \) under metal-free conditions where the formation of \( \text{H}_2 \) is inhibited (Bromm & Loeb 2003), or for gas enriched below the critical metallicity threshold for fragmentation (Santoro & Shull 2006). For these systems, the tenuous gas cools down by atomic hydrogen only until it reaches \( T_{\text{gas}} \sim 4000 \text{ K} \). At this point the cooling function of atomic hydrogen drops by a few orders of magnitude, and contraction proceeds nearly adiabatically. Finally, it has been recently suggested that highly turbulent systems are also likely to experience a limited amount of fragmentation, suggesting that
efficient gas collapse could proceed also in metal-enriched galaxies at later cosmic epochs (Begelman & Shlosman 2009).

In such halos where fragmentation is suppressed and cooling proceeds gradually, the gaseous component can cool and contract until rotational support halts the collapse. In the most common situations, rotational support can halt the collapse before densities required for MBH formation are reached. Additional mechanisms inducing transport of angular momentum are needed to further condense the gas until conditions fostering MBH formation are achieved. An appealing route to efficient angular momentum shedding is by global dynamical instabilities, such as the “bars-within-bars” mechanism (Shlosman et al. 1989; Begelman et al. 2006). A bar can transport angular momentum outward on a dynamical timescale via gravitational and hydrodynamical torques, allowing the radius to shrink. Provided that the gas is able to cool, this shrinkage leads to even greater instability on shorter timescales, and the process cascades. This mechanism is a very attractive candidate for collecting gas in the centers of halos because it works on a dynamical time scale and can operate over many decades of radius.

An alternative way of describing mass inflow is via the Toomre stability parameter, $Q$. When $Q$ approaches unity a disk is subject to gravitational instabilities. If the destabilization of the system is not too violent, instabilities lead to mass infall instead of fragmentation and global star formation (Lodato & Natarajan 2006). This is the case if the inflow rate is below a critical threshold $M_{\text{max}} = 2\alpha_{c}c_{s}^{3}/G$, where the viscosity is $\alpha_{c} \sim 0.12$, the disk is able to be sustained and molecular and metal cooling are not important. Such an unstable disk develops non-axisymmetric spiral structures, which effectively redistribute angular momentum, causing mass inflow. This process stops when the amount of mass transported to the center is enough to make the disk marginally stable.

After gas has efficiently accumulated in the center, what happens next? The gas made available in the central compact region can then form a central massive object. Depending on how fast and efficiently the mass accumulation proceeds, the exact outcome would differ. It is generally thought that VMSs supported by radiation pressure evolve as an $n = 3$ polytrope (Hoyle & Fowler 1963; Baumgarte & Shapiro 1999; Saijo et al. 2002). The fate of a marginally unstable, maximally rotating VMS has been investigated numerically in full general relativity by Shibata & Shapiro (2002). They found that the final object is a Kerr-like black hole (spin parameter $\sim 0.75$) containing 90% of the stellar mass. The fate of an isolated VMS is therefore the formation of a MBH.

If the mass accumulation is fast, however, the outer layers of VMSs are not thermally relaxed during much of the main sequence lifetime of the star (Begelman 2009). Such VMSs can have a complex structure with a convective (polytropic) core surrounded by a convectively stable envelope. After exhausting its hydrogen, the core of a VMS will contract and heat up until it suffers catastrophic neutrino losses and collapses. The initial black hole, with a mass of a few $M_{\odot}$, formed as a result of core-collapse, subsequently grows via accretion from the bloated envelopes that result (“quasistars,” an initially low-mass black hole embedded within a massive, radiation-pressure-supported envelope; see also Begelman et al. 2006). Over time, the black hole grows at the expense of the envelope, until finally the growing luminosity succeeds in unbinding the envelope and the seed MBH is unveiled. The key feature of this scenario is that while the black hole is embedded within the envelope, its growth is limited by the Eddington limit for the whole quasistar, rather than that appropriate for the black hole mass itself. Very rapid growth can then occur at early times, when the envelope mass greatly exceeds the black hole mass.
The masses of the seeds predicted by different models of gas infall and VMS structure vary, but they are typically in the range $M_{\text{BH}} \sim 10^4$–$10^5 \, M_\odot$ (Figure 1).

### 2.3. Stellar-Dynamical Processes

Efficient gas collapse, leading to MBH seed formation, is mutually exclusive with star formation, as competition for the gas supply limits the mass available. Formation of “normal” low-mass stars however, opens up a new scenario for MBH seed formation, if stellar-dynamical rather than gas-dynamical processes are at play. This first episode of efficient star formation can foster the formation of very compact nuclear star clusters (Schneider et al. 2006; Clark et al. 2008; Omukai et al. 2008) where star collisions can lead to the formation of a VMS, possibly leaving a MBH remnant with mass in the range $\sim 10^2$–$10^4 \, M_\odot$. Gaburov et al. (2009) investigate the process of runaway collisions directly with hydrodynamical simulations and show that during collisions, large mass losses are likely, at metallicities larger than $10^{-3}$ solar. Metal-enriched VMSs are therefore expected to end their lives as objects less massive than $\sim 150 \, M_\odot$, either collapsing into MBHs with masses of a few solar masses or exploding as pair-instability supernovae.

The growth of a VMS should be much more efficient at low metallicity. Low metallicity can modify the picture in different ways. First, at sub-solar (but non-zero) metallicity, stars with masses $\gtrsim 40 \, M_\odot$ are thought to collapse directly into a black hole without exploding as supernovae. Second, the mass loss due to winds is much more reduced in metal-poor stars, which greatly helps in increasing the mass of the final remnant.

Devecchi & Volonteri (2009) investigate the formation of MBHs, remnants of VMS formed via stellar collisions in the very first stellar clusters at early cosmic times. The main features of the model are as follows: we consider halos with virial temperatures $T_{\text{vir}} \gtrsim 10^4$ K after the first episode of star formation, hence with a low, but non-zero, metallicity gas content. This set of assumptions ensures that (i) atomic hydrogen cooling can contribute to the gas cooling process, (ii) a UV field has been created by the first stars, and (iii) the gas inside the halo has been mildly polluted by the first metals. The second condition implies that at low density H$_2$ is dissociated and does not contribute to cooling. The third condition ensures that gas can fragment and form low-mass stars only if the gas density is above a certain threshold, $n_{\text{crit,Z}}$ (which depends on the metallicity, see Santoro & Shull 2006): this causes only the highest density regions of a proto-galactic disk to be prone to star formation.

In Toomre-unstable proto-galactic disks, such as those described in the previous section, instabilities lead to mass infall instead of fragmentation into bound clumps and global star formation in the entire disk. The gas inflow increases the central density within a certain compact region $n > n_{\text{crit,Z}}$. Here star formation ensues and a dense star cluster is formed. At metallicities $\sim 10^{-4}$–$10^{-5}$ solar, the typical star cluster masses are of order $10^5 \, M_\odot$ with typical half-mass radii $\sim 1$ pc. Most star clusters therefore go into core collapse in $\lesssim 3$ Myr, and runaway collisions of stars form a VMS (see Shapiro 2004), leading to a MBH remnant.

As the metallicity of the universe increases, the critical density for fragmentation decreases and stars start to form in the entire protogalactic disk. As gas is consumed by star formation, the inflow of gas is no longer efficient, more extended clusters form, and the core-collapse timescale increases until it exceeds the main-sequence lifetime of massive stars. Devecchi & Volonteri (2009) find that typically a fraction $\sim 0.05$ of protogalaxies at $z \sim 10$–20 form black hole seeds, with masses $\sim 1000$–$2000 \, M_\odot$, leading to a mass density in seeds of a few $\sim 10^2 \, M_\odot$ Mpc$^{-3}$. Most of the assumptions of Devecchi & Volonteri (2009) have been conservative, but still the population of seeds is comparable to the case of Population III star remnants discussed, for instance, by Volonteri et al. (2003). The
fraction of high-redshift galaxies seeded with a MBH is about a factor of 10 below the direct collapse case presented by Volonteri et al. (2008), where a seed was assumed to form with a 100% efficiency whenever a protogalaxy disk was Toomre unstable.

3. Observational Signatures of Massive Black Hole Seeds

What are the possible observational tests of MBH formation scenarios? Figure 1 shows three mass functions for three different MBH “seed” models: direct collapse (Volonteri et al. 2008), runaway stellar mergers in high-redshift clusters (Devecchi & Volonteri 2009), and Population III remnants (Volonteri et al. 2003). These are the initial conditions that we would like to probe. In this section, I will focus on discriminants for two different, extreme, scenarios: “light seeds,” derived from Population III remnants, and “heavy seeds,” derived from direct gas collapse.

![Figure 1](image-url)  
**Figure 1.** Mass function of seed MBHs for three different formation scenarios: direct collapse (Volonteri et al. 2008, left), runaway stellar mergers in high-redshift clusters (Devecchi & Volonteri 2009, center), and Population III remnants (Madau & Rees 2001, right). Note the different y-axis scale for the Population III case.

3.1. Tracing MBHs at the Earliest Times

Since during the quasar epoch ($z \approx 3–4$) MBHs increase their mass by a large factor (Yu & Tremaine 2002), signatures of the seed formation mechanisms are likely more evident at earlier epochs. If accretion is efficient, it washes out any trace of the initial seed population. Volonteri & Gnedin (2009) find that of the contribution to the quasar luminosity budget at $z > 6$ is dominated by MBHs with mass $< 10^6 M_\odot$. Such small, low-luminosity MBHs do not contribute to the bright end of the luminosity function of quasars, and are therefore difficult to account for from simple extrapolations of the luminosity function of quasars. These small black holes are not hosted in extremely massive galaxies residing in the highest density peaks (5 to 6σ peaks), but are instead...
found in more common, “normal” systems, $\sim 3\sigma$, peaks. Future generations of space-based telescopes, such as JWST and IXO, are likely to detect and constrain the evolution of the population of accreting massive black holes at early times ($z \gtrsim 10$).

### 3.2. Gravitational Waves

Detection of gravitational waves from seeds merging at the redshift of formation (Sesana et al. 2007) is probably one of the best ways to discriminate among formation mechanisms. LISA in principle is sensitive to gravitational waves from binary MBHs with masses in the range $10^3$–$10^6 M_\odot$ basically at any redshift of interest. A large fraction of coalescences will be directly observable by LISA, and on the basis of the detection rate, constraints can be put on the MBH formation process.

Sesana et al. (2007) and Arun et al. (2009) analyze merger histories of MBHs along the hierarchical build-up of cosmic structures (Volonteri et al. 2003). Sesana et al. (2007) find that a decisive diagnostic is provided by the distribution of the mass ratios in binary coalescences. Models where seeds start large predict that most of the detected events involve equal mass binaries. A fraction of observable coalescences, in fact, involve MBHs at $z > 10$, when MBHs had no time to accrete much mass yet. As most seeds form with similar mass, mergers at early times involve MBH binaries with mass ratio $\sim 1$. In scenarios based on Population III remnants, $z > 10$ mergers involve MBHs with masses below the LISA threshold. The detectable events happen at later times, when MBHs have already experienced a great deal of mass growth, yielding a mass ratio distribution that is flat or features a broad peak at mass ratios $\sim 0.1$–0.2. Arun et al. (2009) determine the detectability of events by using a code that includes both spin precession and higher harmonics in the gravitational-wave signal, carrying out Monte Carlo simulations to determine the number of events that can be detected and accurately localized in these population models. LISA will detect a significant fraction of all mergers: almost all mergers will be detected in heavy-seed scenarios, and nearly half of all mergers in small-seed scenarios. For heavy-seed models, most mergers detectable with high signal-to-noise ratio ($S/N > 10$) occur in the redshift range $3 \lesssim z \lesssim 8$, with a peak around $z \sim 5$. In the case of light seeds, mergers are roughly uniform in $z$ over the range $4 \lesssim z \lesssim 10$.

### 3.3. Massive Black Holes in Low-Mass Galaxies: Occupation Fractions and $M_{\text{BH}} - \sigma$

The repercussions of different initial efficiencies for seed formation for the overall evolution of the MBH population stretch from high-redshift to the local universe. The formation of seeds in a $\Lambda$CDM scenario follows the cosmological bias. As a consequence, the progenitors of massive galaxies (or clusters of galaxies) have a higher probability of hosting MBH seeds (cf. Madau & Rees 2001). In the case of low-bias systems, such as isolated low-mass galaxies, very few of the high-$z$ progenitors have the deep potential wells needed for gas retention and cooling, a prerequisite for MBH formation.

The signature of the efficiency of the formation of MBH seeds will consequently be stronger in low-mass galaxies. Figure 2 (bottom panel) shows a comparison between the observed $M_{\text{BH}} - \sigma$ relation and the one predicted by different models (shown with circles), and in particular, from left to right, two models based on the Lodato & Natarajan (2006) seed masses, and a third model based on lower-mass Population III star seeds. The upper panel of Figure 2 shows the fraction of galaxies that host a massive black hole for different velocity dispersion bins. This shows that the fraction of galaxies hosting an MBH increases with increasing halo masses at $z = 0$. A larger fraction of low-mass halos are devoid of central black holes for lower seed formation efficiencies. This is one of the key discriminants between different “seed” scenarios.
Simple arguments lead us to believe that MBHs might today inhabit the nuclei of dwarf galaxies, such as the satellites of the Milky Way and Andromeda (Van Wassenhove et al. 2009). Indeed, one of the best diagnostics of “seed” formation mechanisms would be to measure the masses of MBHs in dwarf galaxies.  As MBHs grow from lower-mass seeds, it is natural to expect that a leftover population of progenitor MBHs should also exist in the present universe.  As discussed above, the progenitors of massive galaxies have a high probability that the central MBH is not “pristine,” that is, it has increased its mass by accretion, or it has experienced mergers and dynamical interactions.  Any dependence of $M_{\text{BH}}$ on the initial seed mass is largely erased.  Dwarf galaxies undergo a quieter merger history, and as a result, at low masses the MBH occupation fraction and the distribution of MBH masses still retain some “memory” of the original seed mass distribution.  The signatures of MBH formation mechanisms will consequently be stronger in dwarf galaxies (Volonteri et al. 2008).  Van Wassenhove et al. (2009) find that for the most part MBHs hosted in Milky Way satellites retain the original “seed” mass, thus providing a clear indication of what the properties of the seeds were.  MBHs generated as “massive seeds” have larger masses that would favor their identification and their typical occupation fraction is lower, being always below 40% and decreasing to less than 1% for “true” dwarf galaxy sizes.  Population III remnants have a higher occupation fraction, but their masses have not grown much since formation, making their detection harder.  

This can be understood by exploring the establishment and evolution of the empirical correlation between black hole mass ($M_{\text{BH}}$) and velocity dispersion ($\sigma$) as a function of cosmic time in galaxies of different mass (Volonteri & Natarajan 2009).  The massive end of the $M_{\text{BH}}$–$\sigma$ relation is established early, and lower mass MBHs migrate onto it as hierarchical merging proceeds.  The slope and scatter of the relation at the low-mass end instead retains memory of the seeding mechanism.  We find that if MBH seeds are massive, $\sim 10^5 M_\odot$, the low-mass end of the $M_{\text{BH}}$–$\sigma$ flattens towards an asymptotic value, creating a characteristic “plume”.  This plume consists of ungrown seeds that merely continue to track the peak of the seed mass function down to late times.

**Figure 2.** The $M_{\text{BH}}$–$\sigma$ relation at $z = 0$. Every circle represents the central MBH in a halo of given $\sigma$. Observational data are marked by their quoted error bars, both in $\sigma$, and in $M_{\text{BH}}$ (Tremaine et al. 2002). Left to right panels: $Q_c = 1.5, Q_c = 2$, Population III star seeds. Top panels: fraction of galaxies at a given velocity dispersion which do host a central MBH.