# The TERAPIX pipeline

Emmanuel Bertin, Yannick Mellier, Mario Radovich[1], Gilles Missonnier

*Institut d'Astrophysique de Paris, 98bis bd Arago, F-75014 Paris, France*

Pierre Didelon, Bertrand Morin

*CEA/DAPNIA/SAp bat 709, l'Orme les Merisiers, F-91191 Gif-sur-Yvette, France*

**Abstract.** I report on the software development effort involved in the TERAPIX project. TERAPIX is essentially dedicated to the processing of the MEGACAM data; however, TERAPIX software modules are designed for a broader usage, and provide the necessary tools to reduce most CCD imaging surveys. The TERAPIX tasks include automatic pre-reduction, image calibrations, resampling, co-addition and source extraction. I describe our technical choices, as well as the main software features and performances.

## 1. Introduction

During the past decade, the data throughput of astronomical instruments has followed a regular progression, at a rate similar to that of computer performance. Therefore, the development of massive data-processing facilities is still a necessity in projects involving survey instruments.

One of such projects is MEGAPRIME, which consists in the refurbishment of the Canada-France-Hawaii 3.6m telescope's primary focus and the installation of the MEGACAM instrument. TERAPIX (in French: Traitement Elémentaire, Réduction et Analyse des PIXels) is the data processing center associated to MEGAPRIME/MEGACAM. The work done at TERAPIX consists mainly of developing data-processing software, managing computer hardware and operating the pipeline.

This presentation is organized as follows. After a brief description of the MEGACAM instrument and the survey operations (§2.), we expose the goals and the general philosophy of the TERAPIX pipeline (§3.). The individual software modules are described in section §4.. Massive image resampling occurs in the TERAPIX pipeline: section §5. discusses the issues which are associated to it. Ongoing development in image homogenization is presented in §6.. We discuss choices of computer hardware architectures for massive data analysis in §7.. Finally, the current status of the project and its perspectives are summarized in §8..

[1]now at Osservatorio di Capodimonte, via Moiariello 16, 80131 Napoli, Italy

## 2. **The** MEGACAM

The MEGACAM wide-field imager will be installed at the primary focus of the CFH telescope in spring 2002, and will start to operate on a regular basis in fall of the same year. It will observe one full square-degree at a time, in a spectral domain between 0.3 and 1 $\mu$m. MEGACAM is a mosaic CCD camera, designed and built at CEA in Saclay, France (Boulade et al. 2000). As can be seen on Fig. 1, the 40 CCDs do not cover a rectangular area. Each CCD has 2k×4.5k 0.18" pixels, yielding a total of 360 Mpixels per exposure. Despite the large number of pixels, the acquisition of a whole "image" can be done in less than 30 seconds, which makes of MEGACAM an instrument with unprecedented efficiency for both wide and deep sky surveys. As a consequence, one expects up to 15GB of science data per hour during observations, and, typically, 10TB of images per year. For TERAPIX this means that data have to be processed at a sustained rate of 200 kpix/s. As a matter of fact, we aim at 3 times this rate to account for possible hardware and software setbacks.

Figure 1. The mosaic of CCDs seen through the entrance windows of the cryostat of MEGACAM, on its testbench at CEA.

Pre-reduction (bias subtraction, flat-fielding...) and preliminary calibration will be performed on-site using the Elixir pipeline (Magnier et al., this conference). "Almost real-time" detection of supernova and other transient events will also be done at CFHT. The pre-reduced data and their calibration files will be sent for archiving to CADC in Victoria, and to TERAPIX for further processing. Nevertheless, TERAPIX is equipped with a pre-reduction pipeline to allow for re-processing or processing external data.

## 3. TERAPIX: Products and Philosophy

The goal of TERAPIX is to create both "clean", well-calibrated, composited science images, and catalogs of the astronomical sources they contain. A schematic overview of the pipeline is shown in Fig. 2. It consists of software modules controlled by scripts and interfaced to a database.

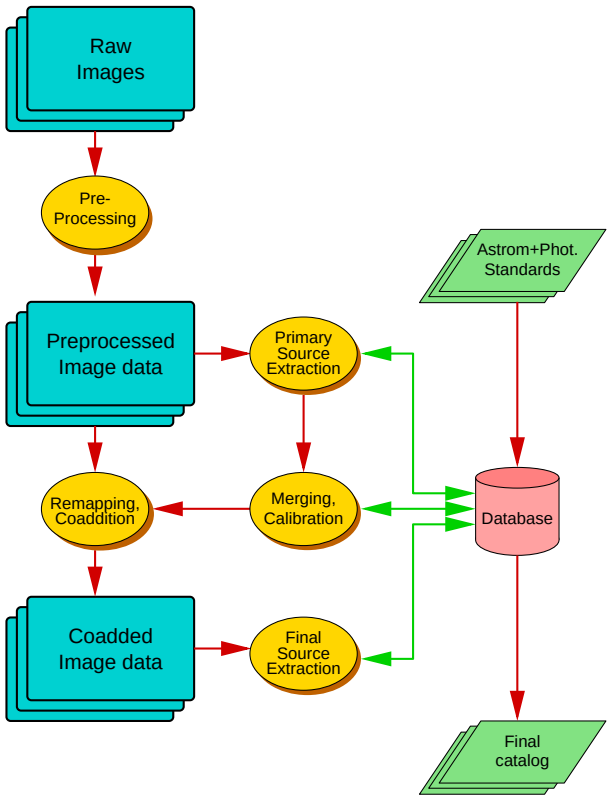


Figure 2. Global overview of the TERAPIX pipeline.

The making of the composite images is by far the most expensive in terms of processing: it involves pixel-weighting, robust source detection, astrometric and photometric calibrations, image resampling and co-addition. For the kind of processing involved in many TERAPIX tasks, input/output bandwidth is as important as sheer computing power. In order to meet the high throughput requirements of the project, it is therefore important to minimize as much as possible exchanges of intermediary image files. By using few pipeline modules that perform many individual tasks, we keep file access to a minimum, and allow the pipeline architecture as seen by the administrator and users to be very simple. Obviously this is possible only because all the modules are developed and maintained in-house. Whenever possible, images are handled as multi-extension FITS files throughout the pipeline. But parallel processing and memory constraints make it necessary to split individual CCD images in some parts of the pipeline.

Originally, the OBJECTIVITY object-oriented database system was selected as our main pipeline database, motivated by the CERN and the SDDS experience. Unfortunately, laborious development, unacceptably low performance on some operations, as well as the recent pessimist claims regarding the future of OBJECTIVITY, convinced us to move back to a more conservative approach: use a relational database. Actually for pipeline work, a simple relational database like MySQL turns out to be performing fast enough on modern hardware, at least as long as detections are not managed individually.

## 4. The TERAPIX software modules

The current list of pipeline software modules and their development status is summarized in the following table (performance is for a 1.3GHz PC):

| Software | Status | Availability to TERAPIX | Availability to users | Performance |
|---|---|---|---|---|
| ASTROMETRIX 1.2 | operational | 2001 | 2002 | 1-10 fields/min. |
| MASKING | in develop. | 2002 | 2003 | - |
| MISSFITS | operational | 2001 | 2002 | 8 Mpixel/s |
| PANORAPIX V.0.9 | operational | 1999 | 2001 | 20 Mpixel/s |
| PHOTOMETRIX 1.0 | operational | 2001 | 2002 | 3-30 fields/min. |
| PREREPIX 1.4 | operational | 2000 | 2002 | 700 kpixel/s |
| PSFEX 1.8.1 | 80% done | 2000 | 2002 | 200 stars/s |
| SEXTRACTOR V.2.2.2 | operational | 1997 | 1995 | 300 sources/s |
| SKYMAKER V.2.3.4 | operational | 1998 | 1998 | 300 stars/s<br>30 galaxies/s |
| STUFF V.1.06 | operational | 1999 | 2000 | 3000 sources/s |
| SWARP V.1.13 | operational | 2000 | 2001 | 200 kpixel/s |
| WEIGHTWATCHER 1.3 | operational | 1997 | 2000 | 2 Mpixel/s |

TERAPIX software modules are released to the public via the Web as soon as we feel that they are stable enough for general use[1]. Although this policy puts more burden on the developer, it has a number of advantages. User feedback brings new suggestions and helps tracing down bugs that appear only in specific configurations. Portability issues can be detected very early. Most of the software packages will have been released before MEGACAM enters in full production.

**MISSFITS.** Upon reception from tape or through FTP, files need to be decompressed, checksummed, and possibly split/joined. This task is devoted to the MISSFITS software. MISSFITS also checks FITS headers conformity and can perform simple operations on FITS keywords (with or without recopy of the original file). MISSFITS is written in C.

[1]Bulletin boards, documentation and download areas can be accessed from `http://terapix.iap.fr/soft`.

WeightWatcher is inherited from the ESO Imaging Survey pipeline. Its task is to combine various masks and gain-maps to prepare the weight-maps of the science images used throughout the pipeline. Weight-maps are essential for processing mosaic data, as they provide estimates of the variance/reliability at each image pixel. They are exploited at various stages of the pipeline, to alter source extractions and optimize image co-additions.

Prerepix is written in Perl/Tk, and makes intensive use of PDL[2] for speed. All the pre-reduction tasks can be configured and scheduled from its GUI: Bias/dark compositing and subtraction, flat-field composition and division, super-flat and fringe composition and correction.

Astrometrix is written in Perl/Tk and C. Astrometrix works in two steps. It first locates automatically the fields on the sky given their approximate position, using a reference catalog (USNO) and a pattern matching algorithm. Once fields are located, overlapping detections among the images are identified and included in the cost function of a global astrometric solution. The cost function, which includes also stars from the reference catalog, is minimized in an iterative way. The iterative approach has the benefit that bringing physical coordinates to a common system is not needed.

Photometrix does for photometry what Astrometrix does for astrometric calibration: homogenizing the zero-points from different pointings to account for different atmospheric extinctions and non-photometric observing conditions.

SExtractor (Bertin & Arnouts 1996) is used at all stages of the pipeline under various configurations for providing source lists. A software module called PSFEx is in charge of deriving automatically a variable, super-sampled model of the Point-Spread-Function. The PSF model is exploited in profile-fitting tasks for stars and galaxies.

SWarp is TERAPIX's image-warping and co-addition tool. As outlined in Fig. 3, it operates on the pre-reduced images and their weight-maps. Based on the astrometric and photometric calibrations derived at an earlier phase of the pipeline, SWarp re-maps ("warps") the pixels to a perfect projection system, and co-adds them in an optimum way, according to their relative weights. SWarp's astrometric engine is based on a customized version of M.Calabretta's WCSLib 2.6 and supports all the projections defined in the 2000's version of the WCS proposal (Greisen & Calabretta 2000), as shown in Fig. 4.

For imaging programs that survey a large angular area (several degrees in width), an output equal area projection is to be preferred to avoid having to deal with a variable pixel scale.

[2] `http://pdl.perl.org/`

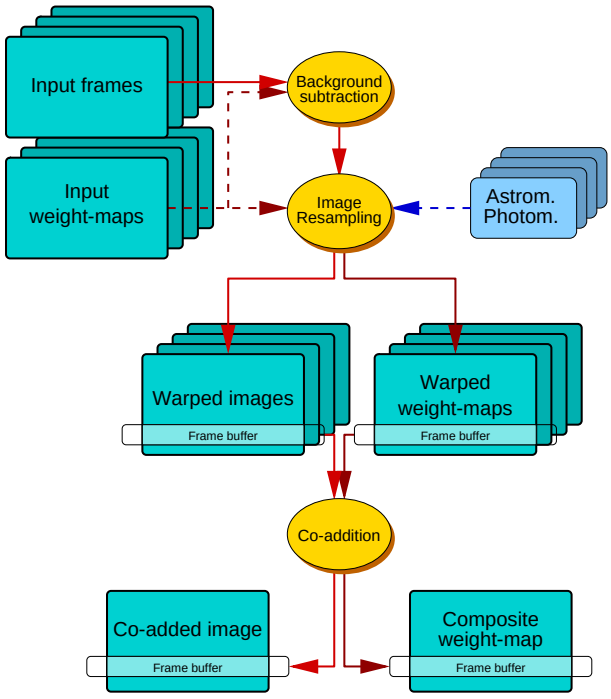


Figure 3. Global overview of the SWarp package.

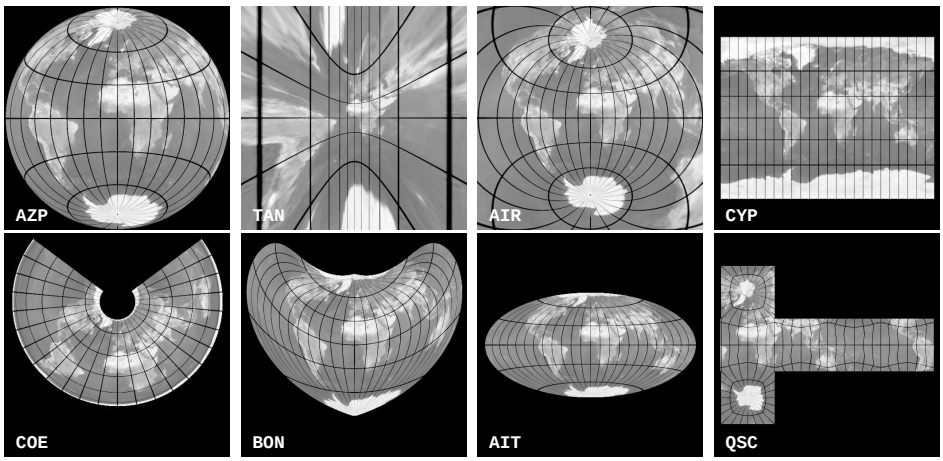


Figure 4. Some projections performed by the SWarp package.

## 5. Image resampling

A critical issue associated with image warping is resampling. Image resampling has a reputation for "corrupting" pixel values sufficiently so that it may affect the scientific content of images. Well-known consequences include the degradation of bright star photometry, or moiré effects on the background noise. Linear resampling on a regular grid is dictated by the choice of an interpolation function. For properly sampled images like those of MEGACAM (Full-Width at Half-Maximum $\geq 3$ pixels), one can afford interpolation functions with negative side-lobes. The best compromise between image fidelity and "localizability" of the interpolation was found for interpolation kernels $6 \times 6$ pixels in size, like the LANCZOS3 kernel. As illustrated in Fig. 5, interpolation of properly sampled data with this kernel makes the effects of resampling on photometric and astrometric measurements negligible.

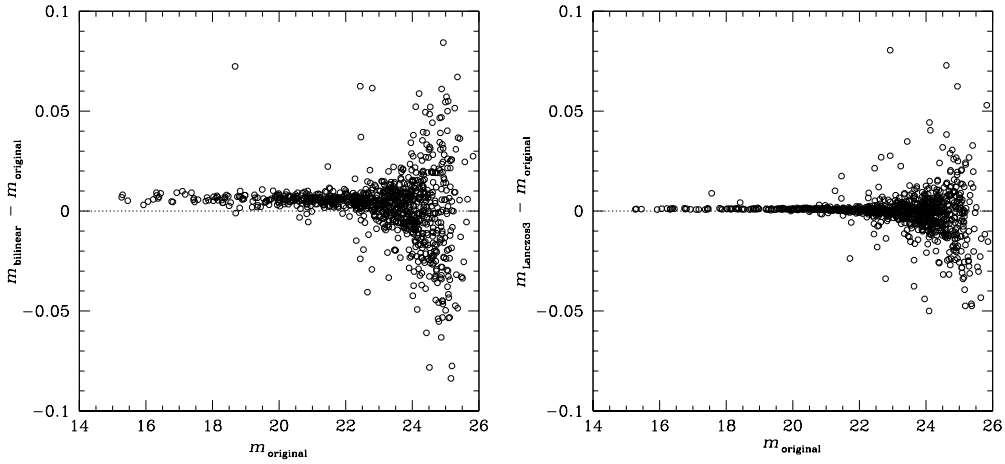


Figure 5. Effects of the resampling on flux measurements. *Left*: bilinear interpolation; *right*: Lanczos3 interpolation. In both cases, a simulated deep sky image with 0.7" seeing, containing stars and white background noise, was rotated by 20 degrees and then rotated back to match the original image. Fluxes were measured in a fixed 2" aperture. The dispersions seen here reflect the differences between *measurements* on the original and resampled images. These dispersions are much smaller than what one would observe by comparing the measurements on the resampled images with the theoretical (noise-free) fluxes of the simulation. Note the significant dispersion in magnitude for the bilinear case, consequences of the stronger smoothing induced by bilinear interpolation.

## 6. Image homogenization

Temporal variability of the seeing is a major problem when compositing dithered images from ground-based instruments. Although the smooth spatial variations of the PSF can easily be modeled on individual images, variations from exposure to exposure — mostly due to seeing fluctuations — will lead to abrupt changes at the frame boundaries in the final combined image. These sharp transitions in image quality from zone to zone cannot easily be modeled, and have a severe impact on the homogeneity of profile-fitting astrometry and photometry, star/galaxy separation and shape measurements. Furthermore, PSF variability does also seriously compromise the effectiveness of non-linear co-addition schemes like median and panchromatic "$\chi^2$" (Szalay et al. 1999) combinations. Hence there is no doubt that homogenizing the PSF across exposures is a necessity for forthcoming large imaging surveys. PSF-homogenization is now commonplace in microlensing experiments for which efficient image subtraction techniques have been developed (e.g. Alard & Lupton 1998). For co-addition we have the additional constraint that a PSF must be imposed: a sensible choice is an isotropic PSF with the average FWHM of the survey. A version of SWarp implementing PSF-homogenization is currently being developed. Homogenized versions of two simulated images with different seeings are shown in Fig. 6. Experience shows that seeing variations of about 50% can perfectly be accommodated with current methods. A large fraction of observations lies within this range in good astronomical sites, which makes image homogenization worthwhile for many surveys.

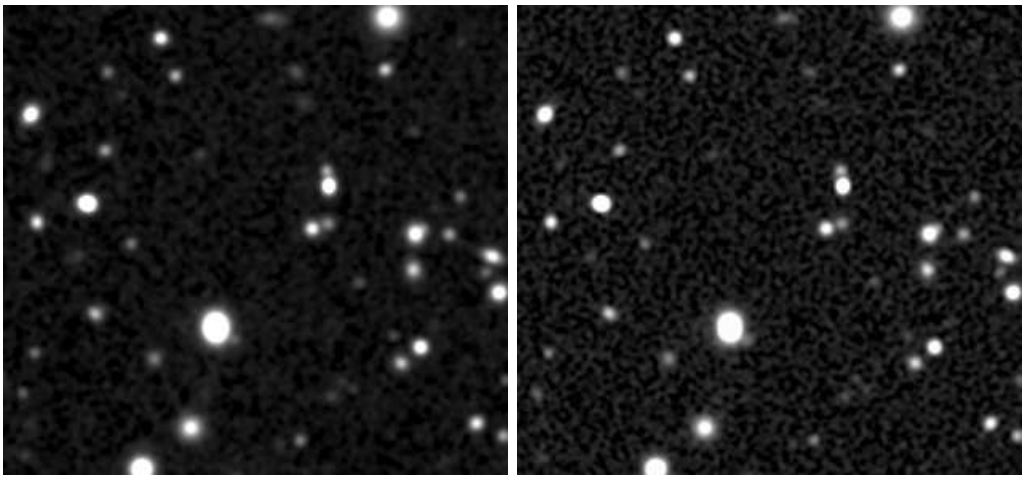

Figure 6. Test of PSF homogenization on simulated images. *Left*: MEGACAM image with an original seeing of 0.6", degraded to 0.75"; *right*: MEGACAM image of the same field (same exposure time) with an original seeing of 0.9", sharpened to 0.75".

Convolving the images with a variable homogenization kernel correlates the photon and instrumental noises at the PSF scale, and it does it in a variable way: somehow, we just shifted the PSF problem elsewhere. Fortunately, when co-adding a large number of images the net effects of high-pass and low-pass

filtering of individual exposures often compensate quite well. In the worst case (all the images in a part of the surveyed area having a bad or a good seeing), a noise autocorrelation map shall be derived prior to sensitive analyses.

## 7. TERAPIX hardware

Up to now (end of 2001), TERAPIX data reduction has been running on Alpha workstations. We are migrating at present to a cluster of Linux dual-processor PCs[3] (Fig.7), which offer unbeatable performance-to-price ratios. PC microprocessors have become so powerful that the limitation in data processing efficiency now mostly comes from input/output bandwidth, especially when machines are run in parallel. Like the Elixir pipeline, TERAPIX has made the choice of not using explicitely parallel code, but instead to run tasks in parallel on the different machines. Although unsophisticated, this approach has many advantages: no additional software development is needed; it makes the pipeline both flexible and portable; data transfers between the computing units are minimized and network latencies do not affect the efficiency of the whole processing. One remaining handicap of current PCs is their limitation to 32 bit addressing, which in practice with Linux does not allow more than $\approx$ 3GB of memory. This is indeed a serious limitation for processing images from large CCD cameras (one single MEGACAM exposure weighs in at 1.4GB in floats). Hopefully the next generation of 64 bit PCs that should become available in early 2003 will make the manipulation of very large images in one piece possible.

## 8. Current status and perspectives

The pipeline is already operational in a rudimentary form with Perl scripts and current software modules. So far, a total of more than 50 square degrees of final scientific images have been produced from several wide-field instruments (UH8k, CFH12k, WFI) in the context of the VIRMOS photometric survey (McCracken et al. 2002) and other smaller private programs. Scientific validation on particularly sensitive aspects (weak gravitational lensing, galaxy correlation functions) is being conducted in parallel.

Several important issues remain to be solved before the start of MEGACAM in fall 2002, in particular the automatic identification and removal of optical ghosts that plague deep observations done with wide-field instruments. The European projects that were launched in 2001 (AVO[4], Astro-WISE[5], and DataGrid[6]), in which TERAPIX is involved, bring cooperations between datacenters within EU to a new level. This shall accelerate the resolution of technical

[3]Informations about the TERAPIX PC cluster, configuration and miscellaneous hardware tips can be found at `http://terapix.iap.fr/hard/`.

[4]`http://www.eso.org/projects/avo/`

[5]`http://www.astro-wise.org/`

[6]`http://www.eu-datagrid.org/`

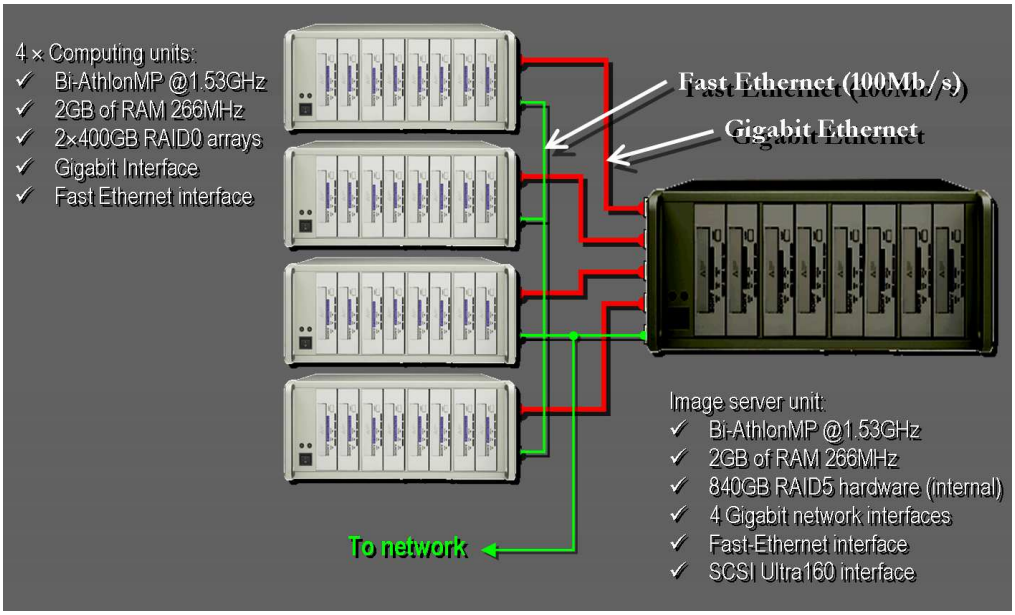


Figure 7. Overview of the cluster of data-processing PCs at TERAPIX. The disk arrays mentioned here are used as "scratch disks" for processing.

problems on our way to (why not?) a universal pipeline and a link between the real and the virtual observatories.

## References

Alard C., Lupton R.H., 1998, ApJ, 503, 325

Bertin E., Arnouts S., 1996, A&AS, 117, 393

Boulade O., Charlot X., Abbon P. et al., 2000, in SPIE Proc., Vol. 4008, 657

Calabretta M.R., Greisen E.W., 2000, to be submitted to A&A

McCracken H.-J.., Radovich M., Foucaud S., Bertin E., Mellier Y., Dantel-Fort M., le Fèvre O., 2002, in preparation

Szalay A.S., Connolly A.J., Szokoly G.P., 1999, AJ, 117, 68