Efficient data reduction and analysis of DECam images using multicore architecture
Poor man's approach to Big data

Roberto Muñoz

Instituto de Astrofísica
Pontificia Universidad Católica de Chile

Thomas Puzia, Maren Hempel, Paul Eigenthaler,
Matt Taylor, Yasna Ordenes, Simón Angel

Ariane Lançon, Steffen Mieske, Michael Hilker,
Patrick Côté, Laura Ferrarese & NGFS team.

Tools for Astronomical Big Data
Tucson, Arizona 2015

Challenges of Big Data

- Computer scientists have been aware of big data for many years, maybe decades. They paved and keep paving the road for dealing with big amounts of data.
- Three concepts are repeated: Volume, Variety and Velocity

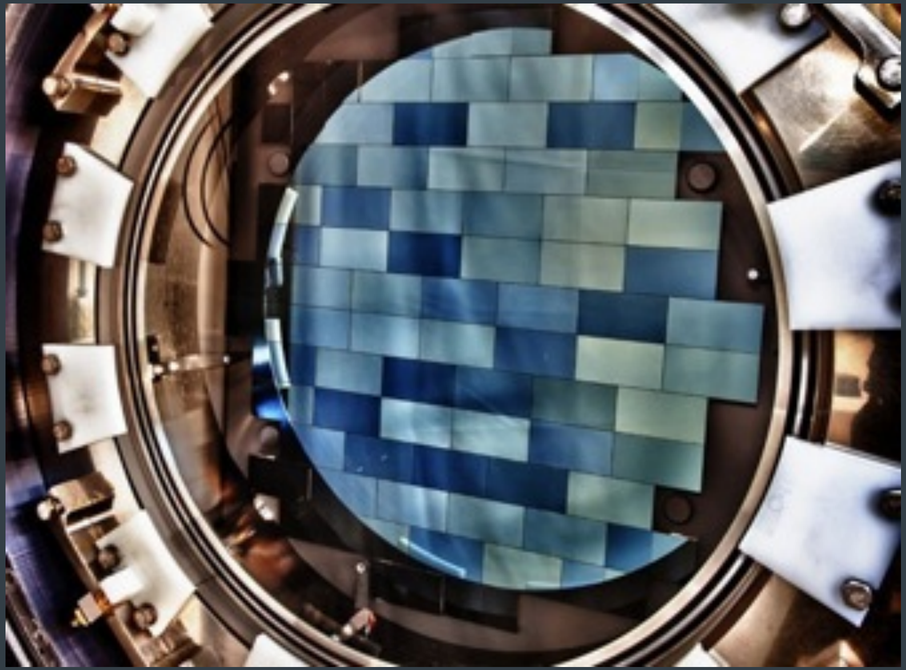
Volume: Quantity of data. GB, TB, PT.
How large will be the data release?
Variety: Different types of data.
Infrared images, High-res spectra
Velocity: Rate at which generate data.
GB, TB or PT per night?

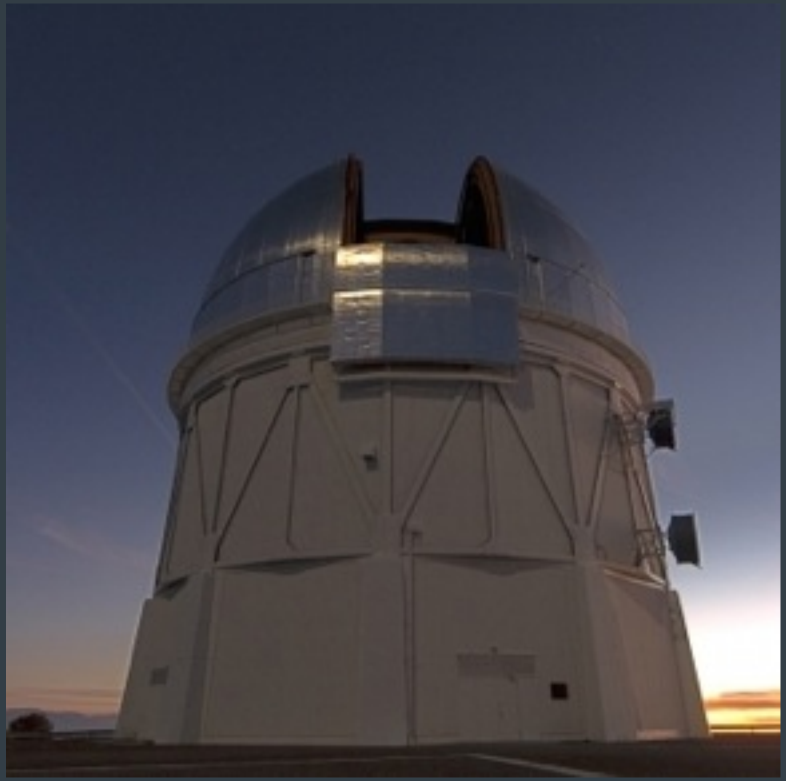
Astronomy and big data

- Science is the main driver of any scientific project where astronomers are involved.
- We were not looking for big amounts of data in the Universe, it just happened that cameras became bigger and bigger and the data processing became more complex.
- Maybe the first project where astronomers were forced to face big data challenges was the SDSS survey. Final data release (DR12) consists of 100 TB.

What about DECam?

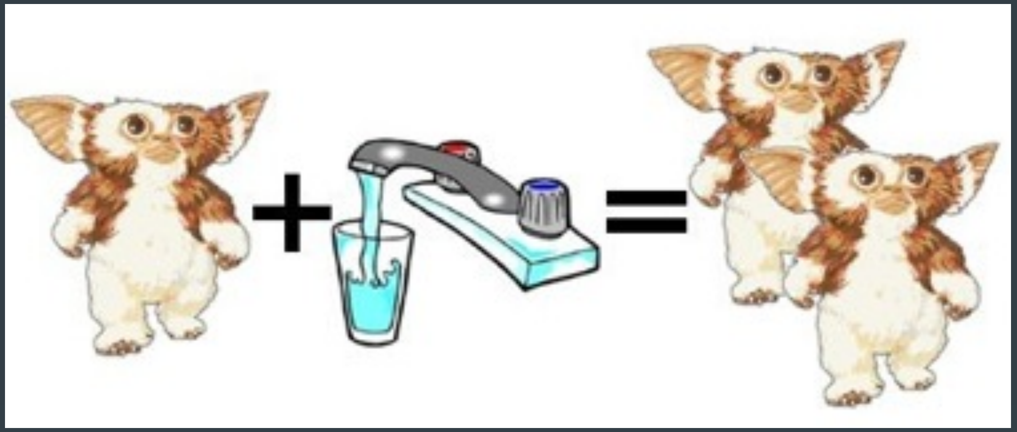
- Each image is about 1 GB and raw data per night could be up to 0,5 TB. At this stage 16 bit signed integer.
- All the raw data is transferred to NOAO servers, processed by the Community pipeline and then the PI is able to download different types of products. At this stage 32 bit floating point.
- Just downloading the images and weight maps, the 0,5 TB/night becomes 3 TB/night of just Community pipeline processed data.





What about post processing?

- Depending on the science you want to do with DECam, maybe you need to download all the individual images processed by the Community pipeline and do some post processing.
- Let’s say you have a very particular observational strategy and want to model the sky from each individual image. But for doing that, you need to know what pixels belong to the sky.
- The easiest way is building a first stack, detecting the sources (SExtractor) and masking them. You end up with a mask per image.
- Since you can’t load the entire 3 TB of data into memory, you need to create the individual sky and sky subtracted images, and then you can stack again.



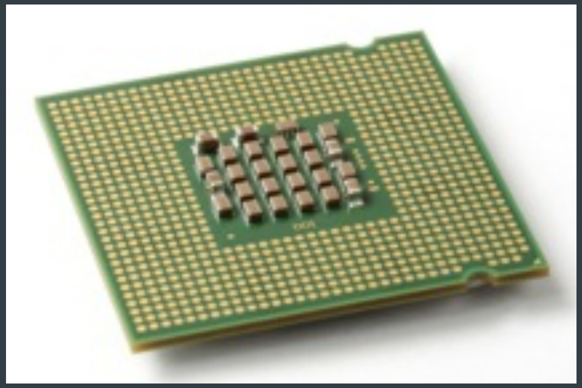
0,5 TB → 5 TB
per night

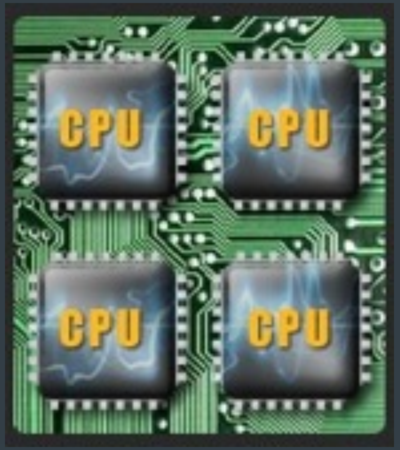
RAID systems

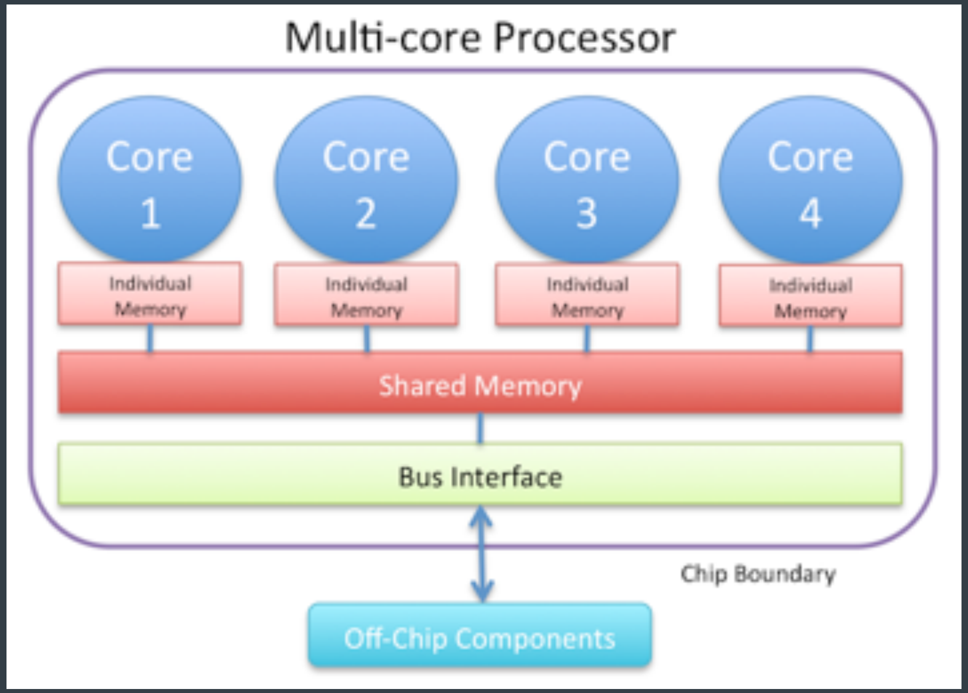
- Our programs consists of about 5 nights per year. We need about 25 TB of space for post processing.
- We found that video professionals (movies) have been using small and medium sizes RAID systems for a while. Nowadays, the technology is cheaper and small groups can setup a system.

Multicore systems

- Multicore systems started becoming popular around 2005. Most of the computers and smartphones we use nowadays are multicore.
- Large storage capacity + High I/O data transfer + Fast CPU processing + Low cost = Short execution times and accessibility for testing new algorithms for image processing

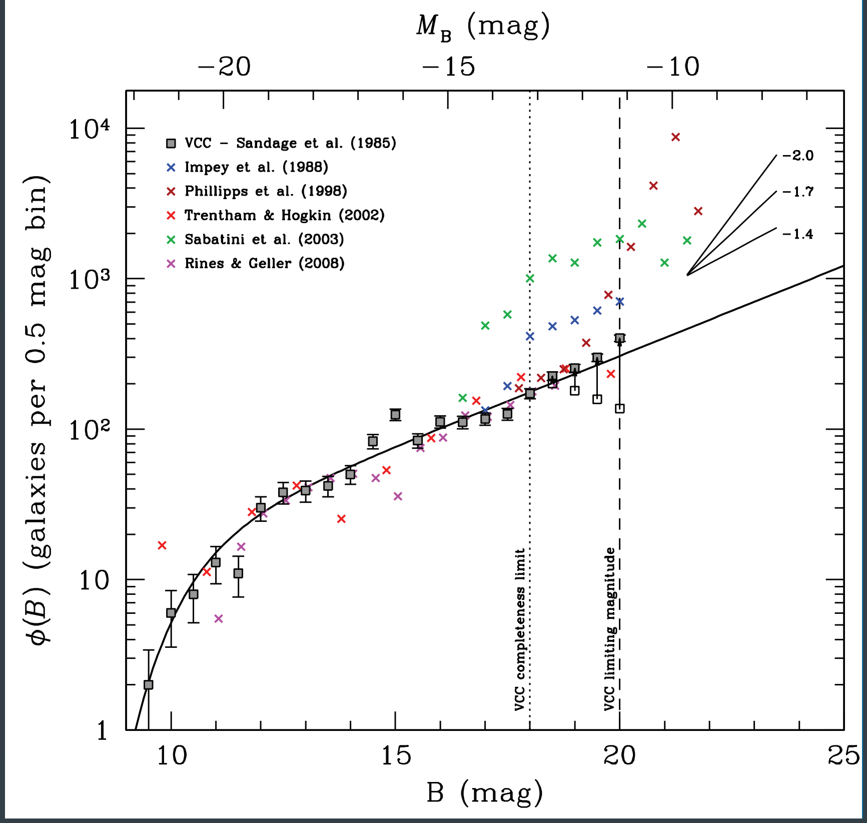


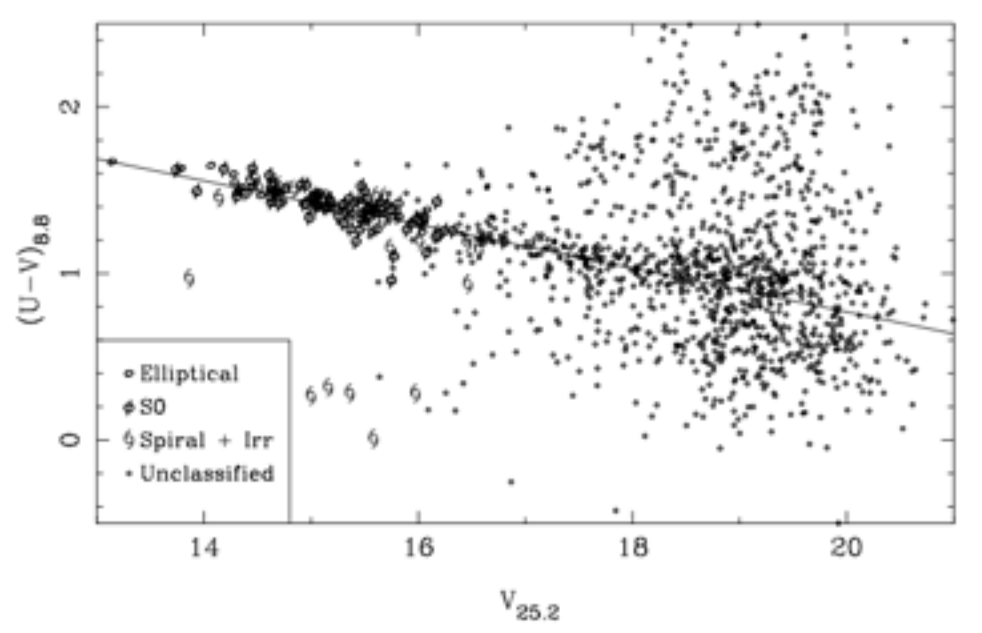




Science: groups and clusters

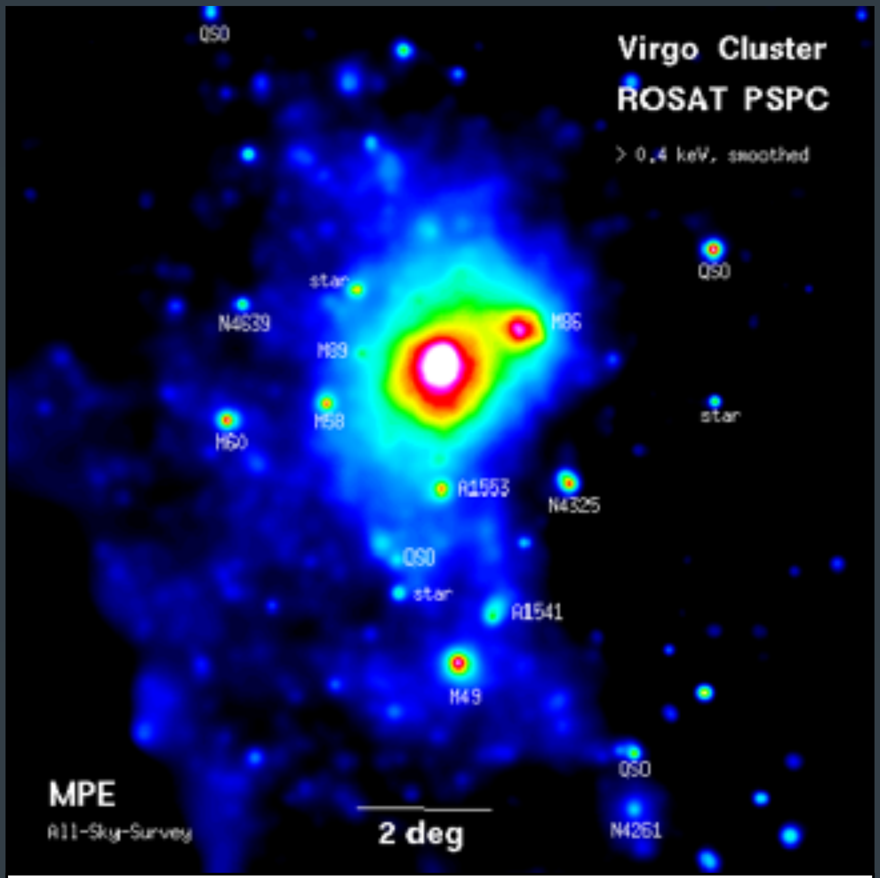
- Our team is using DECam to observe several galaxy groups and clusters closer than 20 Mpc.
- We are interested in studying the stellar mass function down to 10^6 M_O, how environment affects galaxy evolution and how the GCs are distributed around galaxies and ICM, among several other goals.
- Similar questions were asked by Binggeli et al. (1985) about the Virgo galaxy cluster and by Ferguson et al. (1989) about the Fornax galaxy cluster.





NGVS: Virgo cluster

- The Next Generation Virgo Survey (NGVS; PI: L. Ferrarese) is designed to provide deep, high spatial resolution, and contiguous coverage of the Virgo cluster from its core to virial radius.

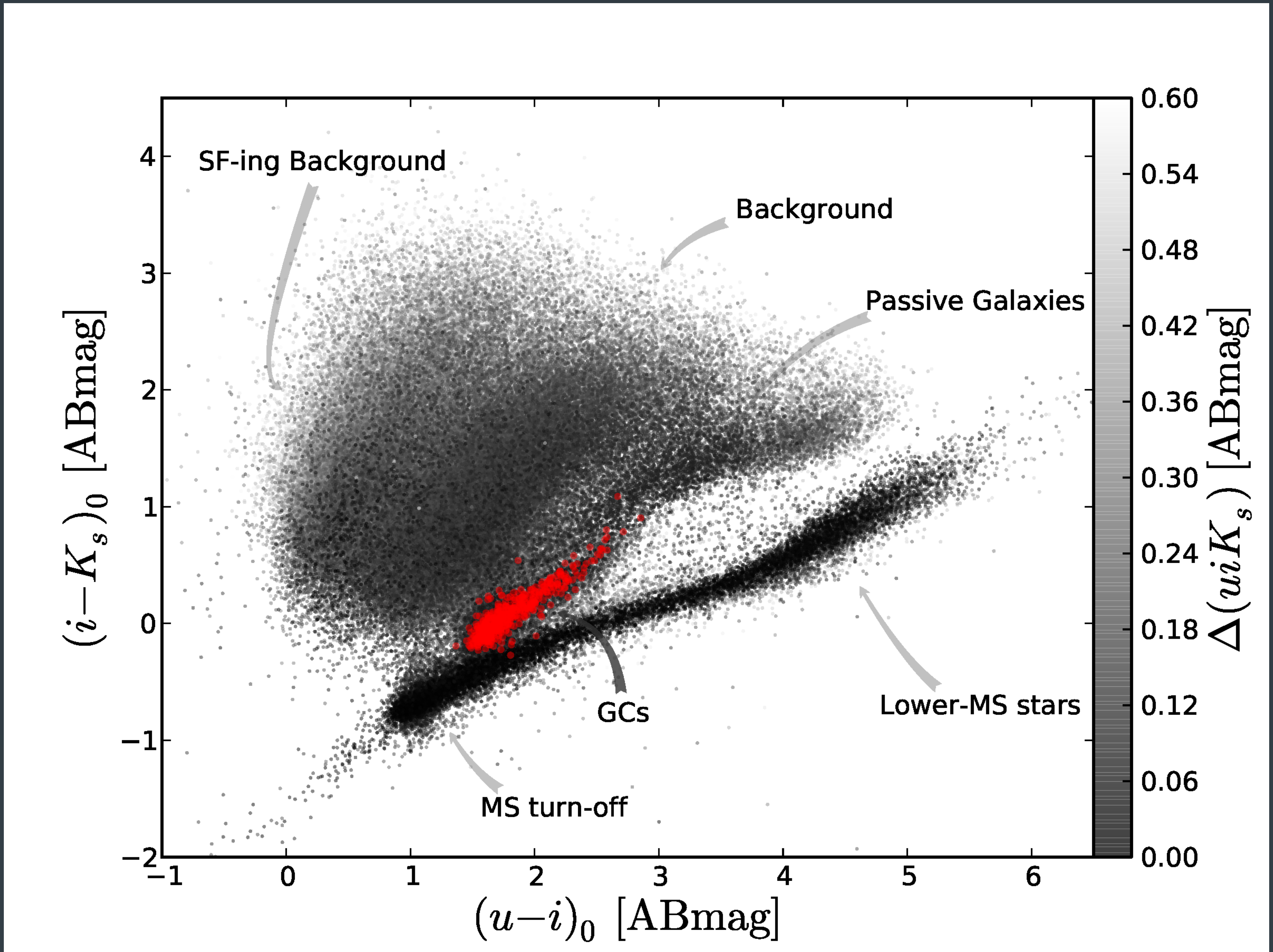


Virgo cluster in X-ray from ROSAT (Böhringer et al. 1994)

The Next Generation Virgo Cluster Survey

The NGVS as it would appear in the sky
Photo Jean-Charles Cuillandre (2008)

The uiK diagram

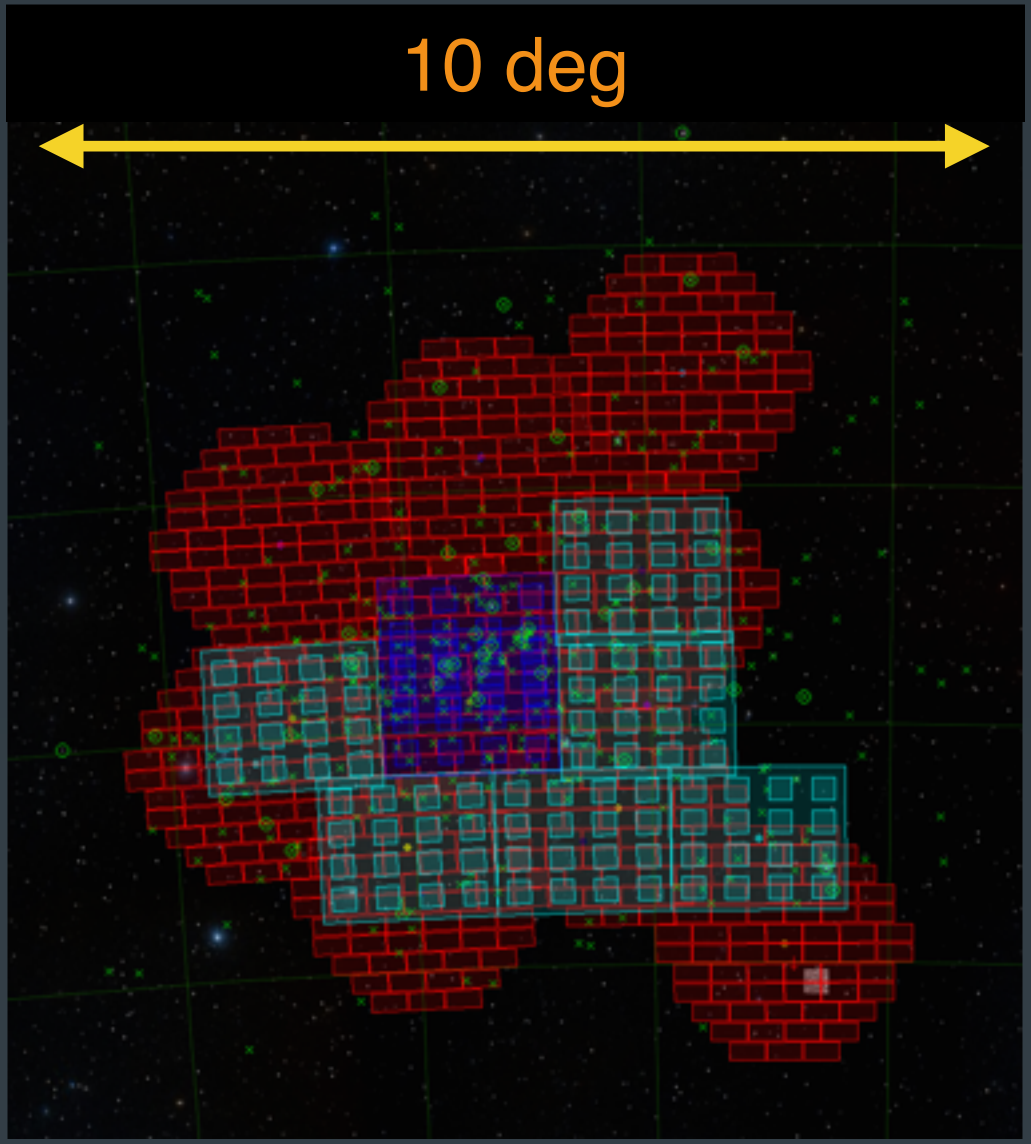


Muñoz et al. 2014

NGFS: Fornax

- The Next Generation Fornax Survey (NGFS; PI: R. Muñoz) is an ongoing multipassband optical and NIR survey of the Fornax galaxy cluster. It will cover the central 30 deg^2 out to the virial radius and will allow to study the galaxy and GC populations.

Astrophotography by Marzo Lorenzini

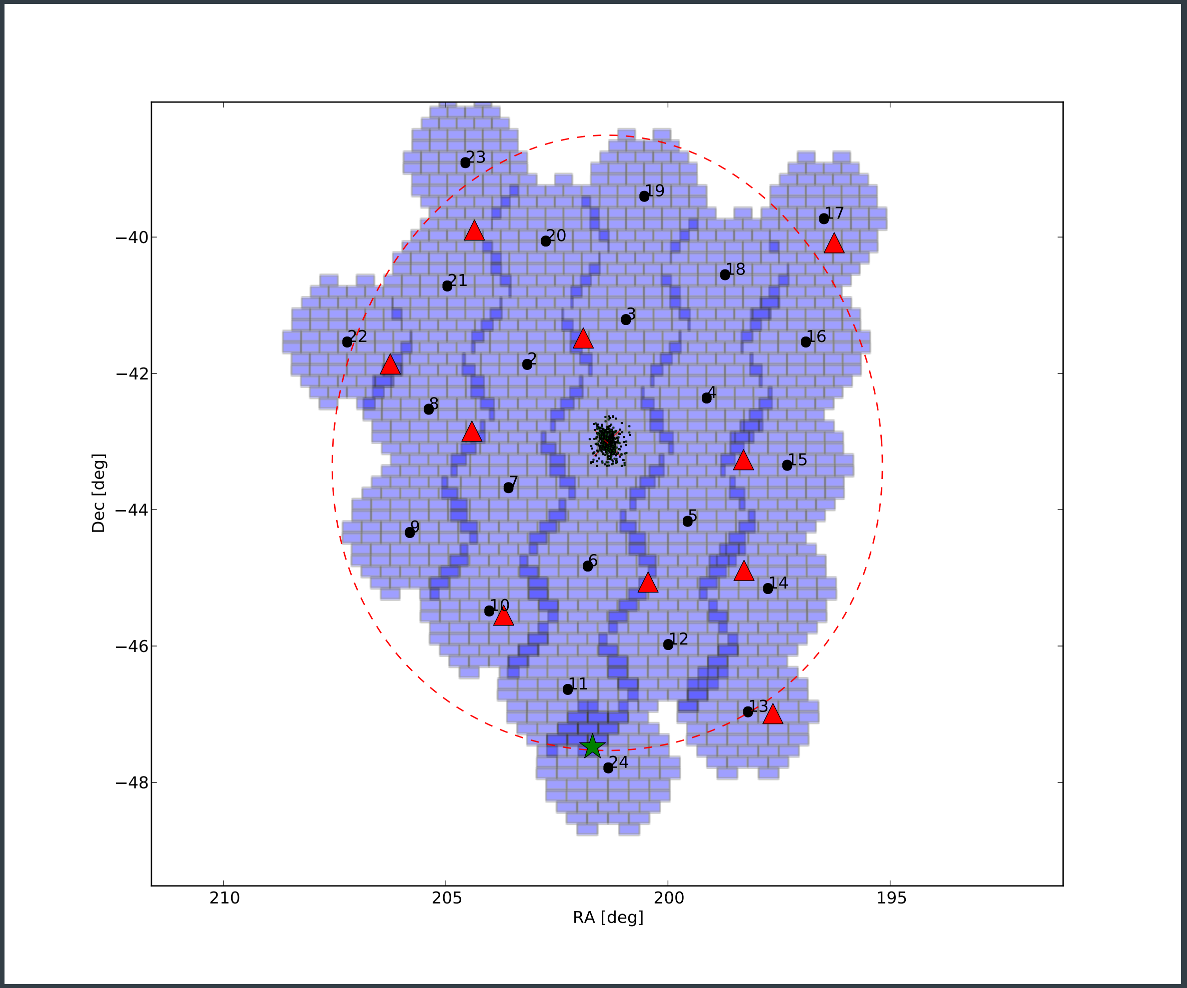


SCABS: Centaurus A

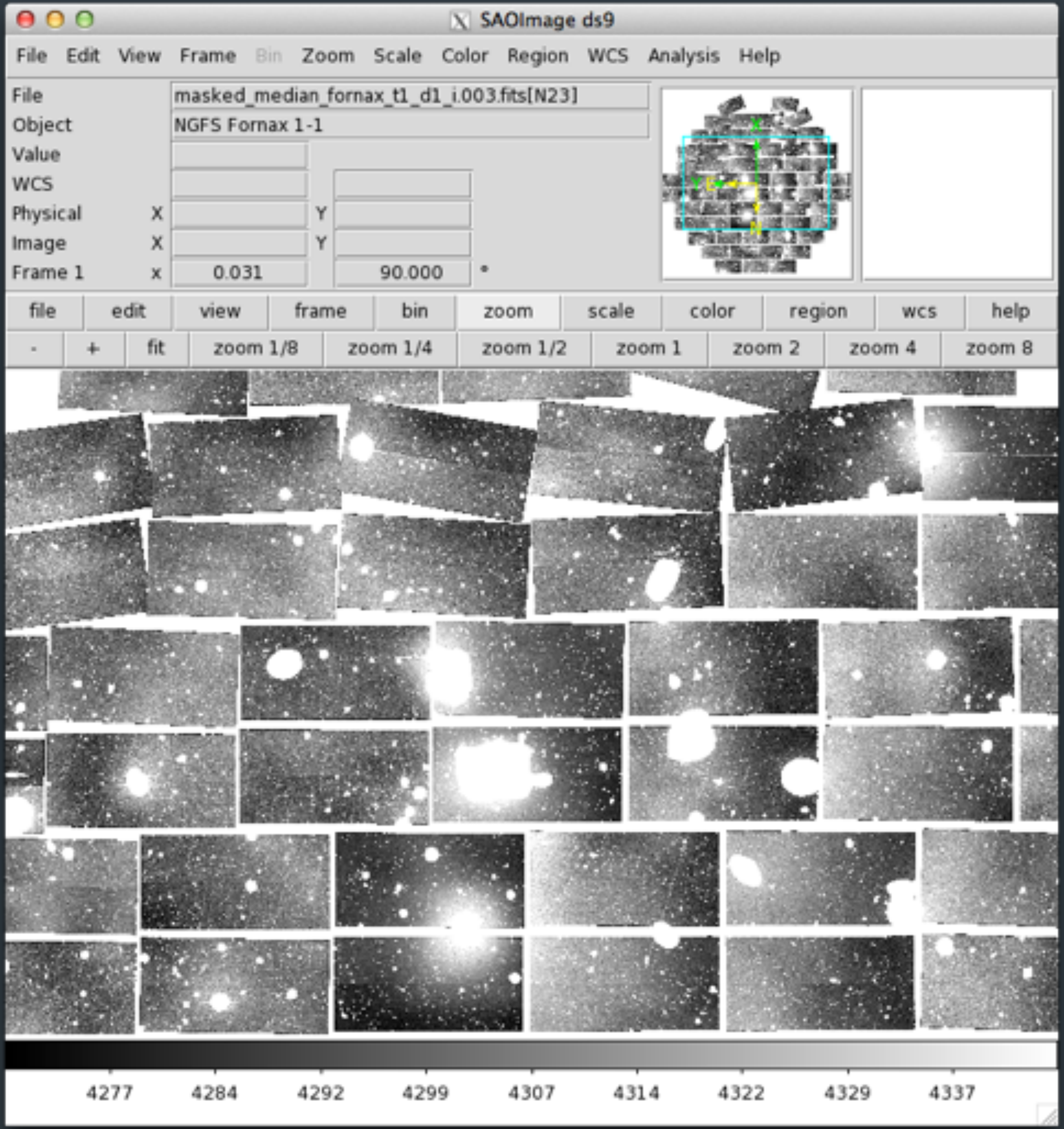
- The Survey of Centaurus A's Baryonic Structures (SCABS; PI: M Taylor) is mapping 72 deg^2 around Cen A.

7 arcmin

20 deg



Sky subtraction: Polynomia

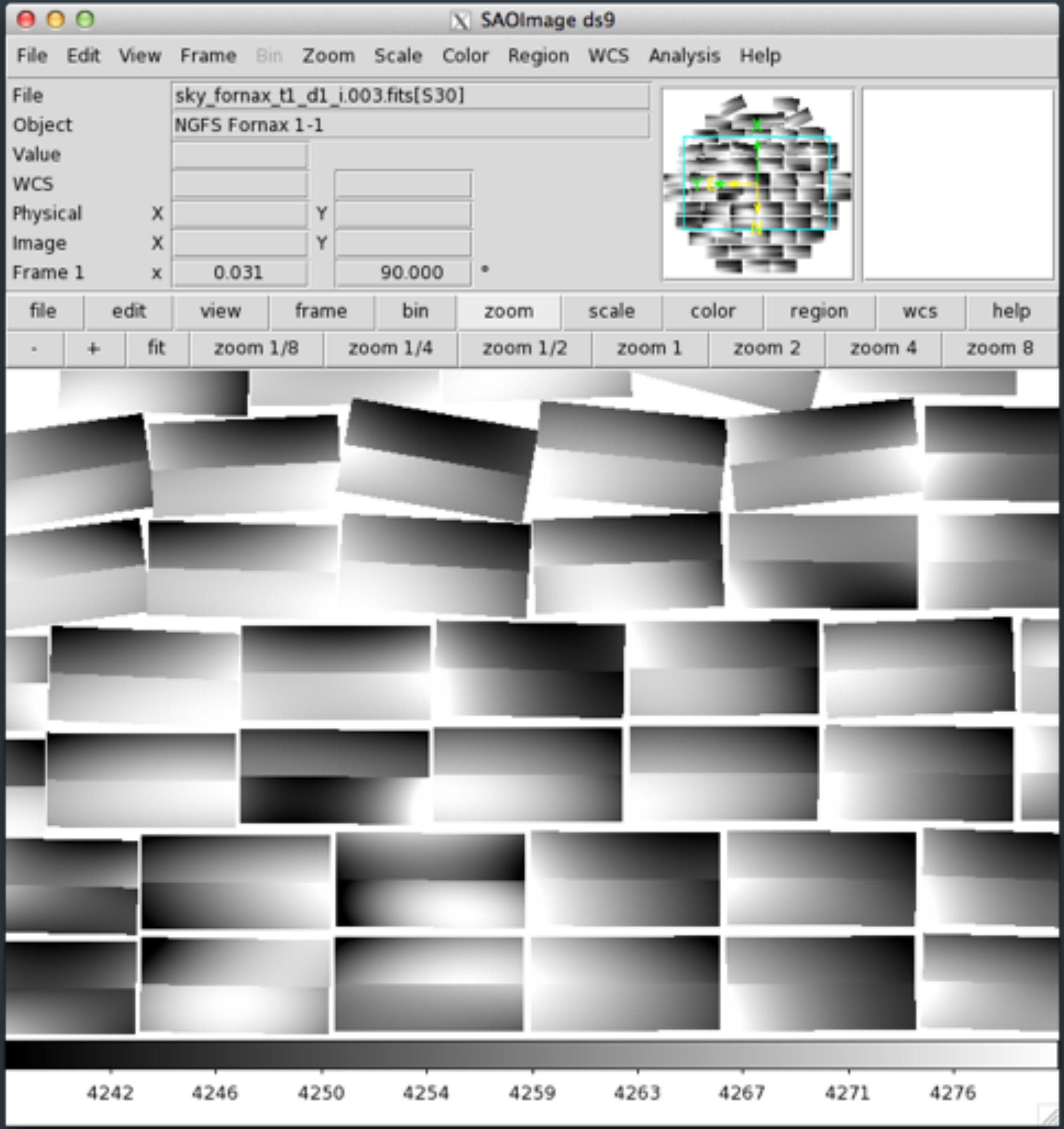


File masked_median_fornax_t1_d1_i.003.fits[N23]

Object NGFS Fornax 1-1

Frame 1 x 0.031 90.000 °

SAOImage ds9
File Edit View Frame Bin Zoom Scale Color Region WCS Analysis Help
Value
WCS
Physical X Y
Image X Y
file edit view frame bin zoom scale color region wcs help
- + fit zoom 1/8 zoom 1/4 zoom 1/2 zoom 1 zoom 2 zoom 4 zoom 8
4277 4284 4292 4299 4307 4314 4322 4329 4337



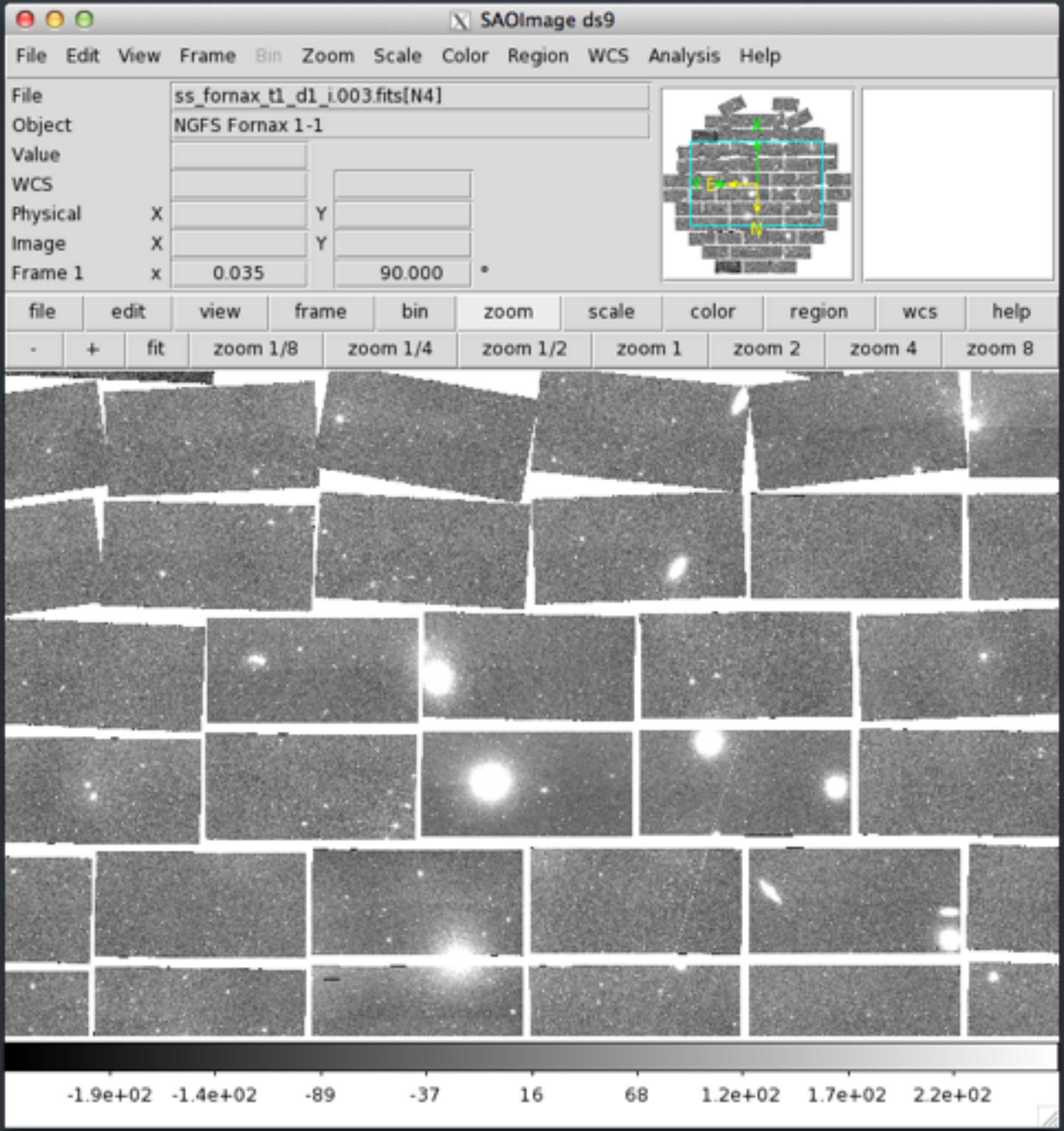
File sky_fornax_t1_d1_i.003.fits[S30]

Object NGFS Fornax 1-1

Frame 1 x 0.031 90.000 °

SAOImage ds9
File Edit View Frame Bin Zoom Scale Color Region WCS Analysis Help
Value
WCS
Physical X Y
Image X Y
file edit view frame bin zoom scale color region wcs help
- + fit zoom 1/8 zoom 1/4 zoom 1/2 zoom 1 zoom 2 zoom 4 zoom 8
4242 4246 4250 4254 4259 4263 4267 4271 4276

Sky subtraction: Polynomia



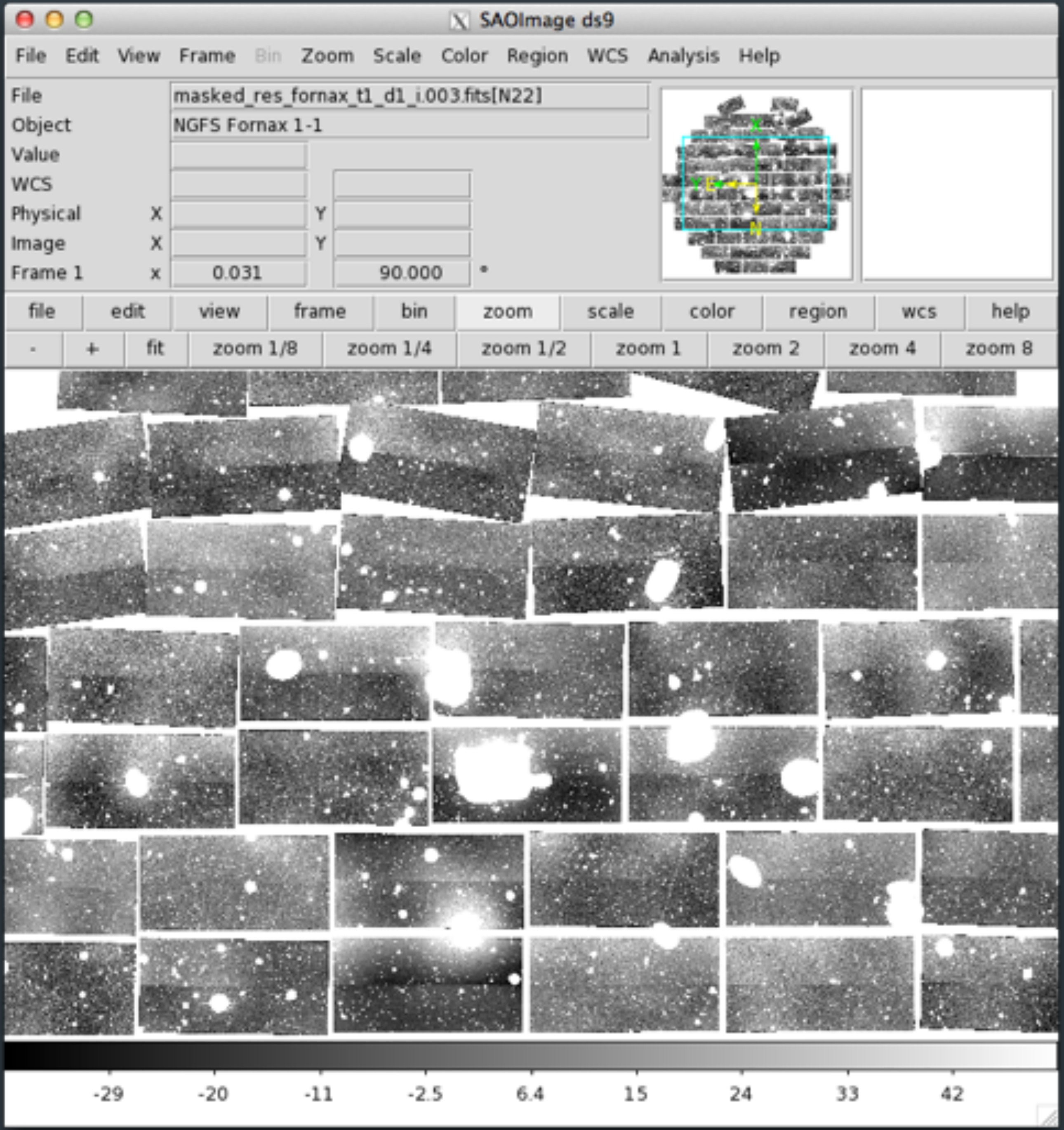
SAOImage ds9

File ss_fornax_t1_d1_i.003.fits[N4]

Object NGFS Fornax 1-1

Frame 1 x 0.035 90.000 °

File Edit View Frame Bin Zoom Scale Color Region WCS Analysis Help
Value
WCS
Physical X Y
Image X Y
file edit view frame bin zoom scale color region wcs help
- + fit zoom 1/8 zoom 1/4 zoom 1/2 zoom 1 zoom 2 zoom 4 zoom 8
-1.9e+02 -1.4e+02 -89 -37 16 68 1.2e+02 1.7e+02 2.2e+02



SAOImage ds9

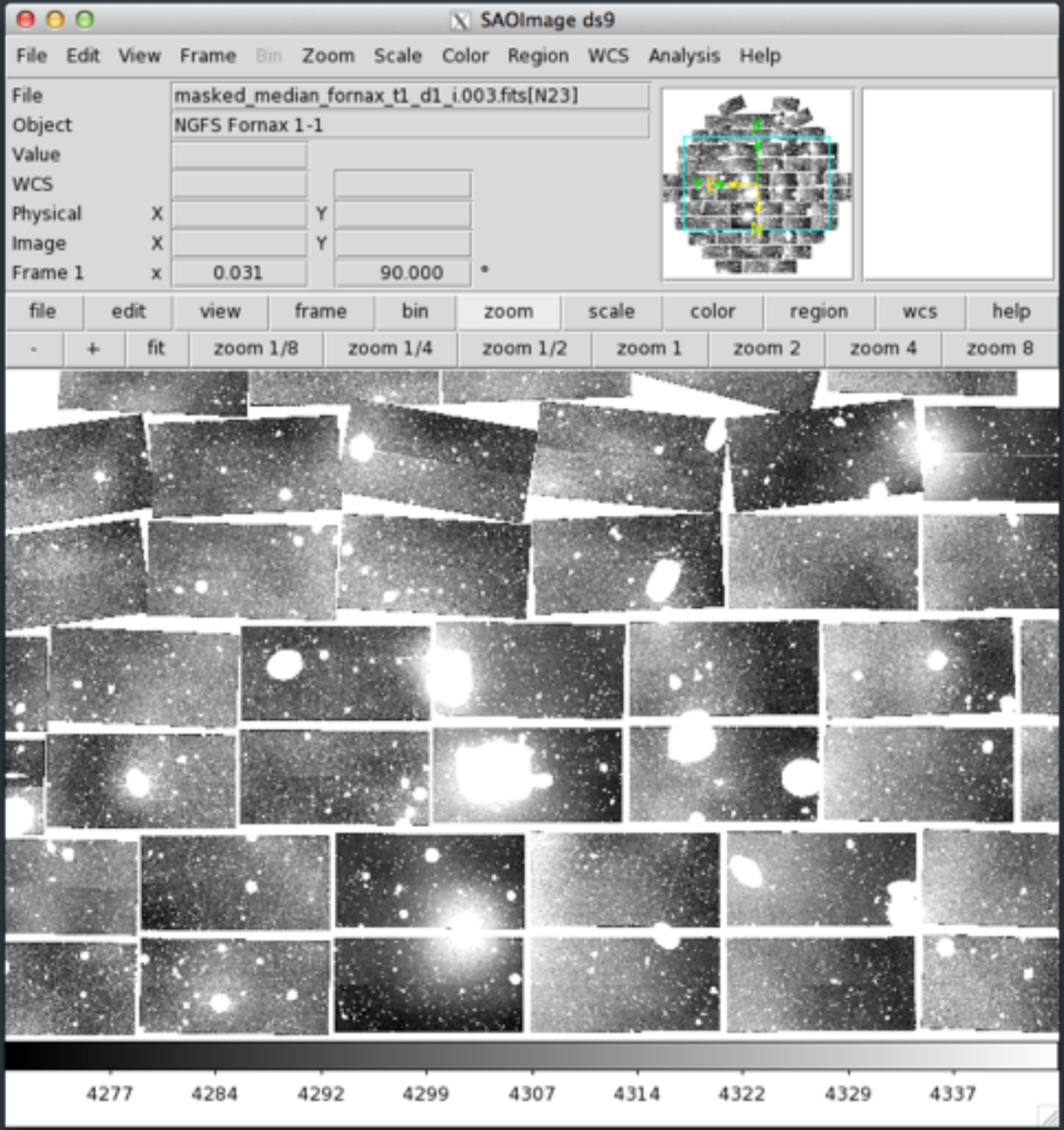
File masked_res_fornax_t1_d1_i.003.fits[N22]

Object NGFS Fornax 1-1

Frame 1 x 0.031 90.000 °

File Edit View Frame Bin Zoom Scale Color Region WCS Analysis Help
Value
WCS
Physical X Y
Image X Y
file edit view frame bin zoom scale color region wcs help
- + fit zoom 1/8 zoom 1/4 zoom 1/2 zoom 1 zoom 2 zoom 4 zoom 8
-29 -20 -11 -2.5 6.4 15 24 33 42

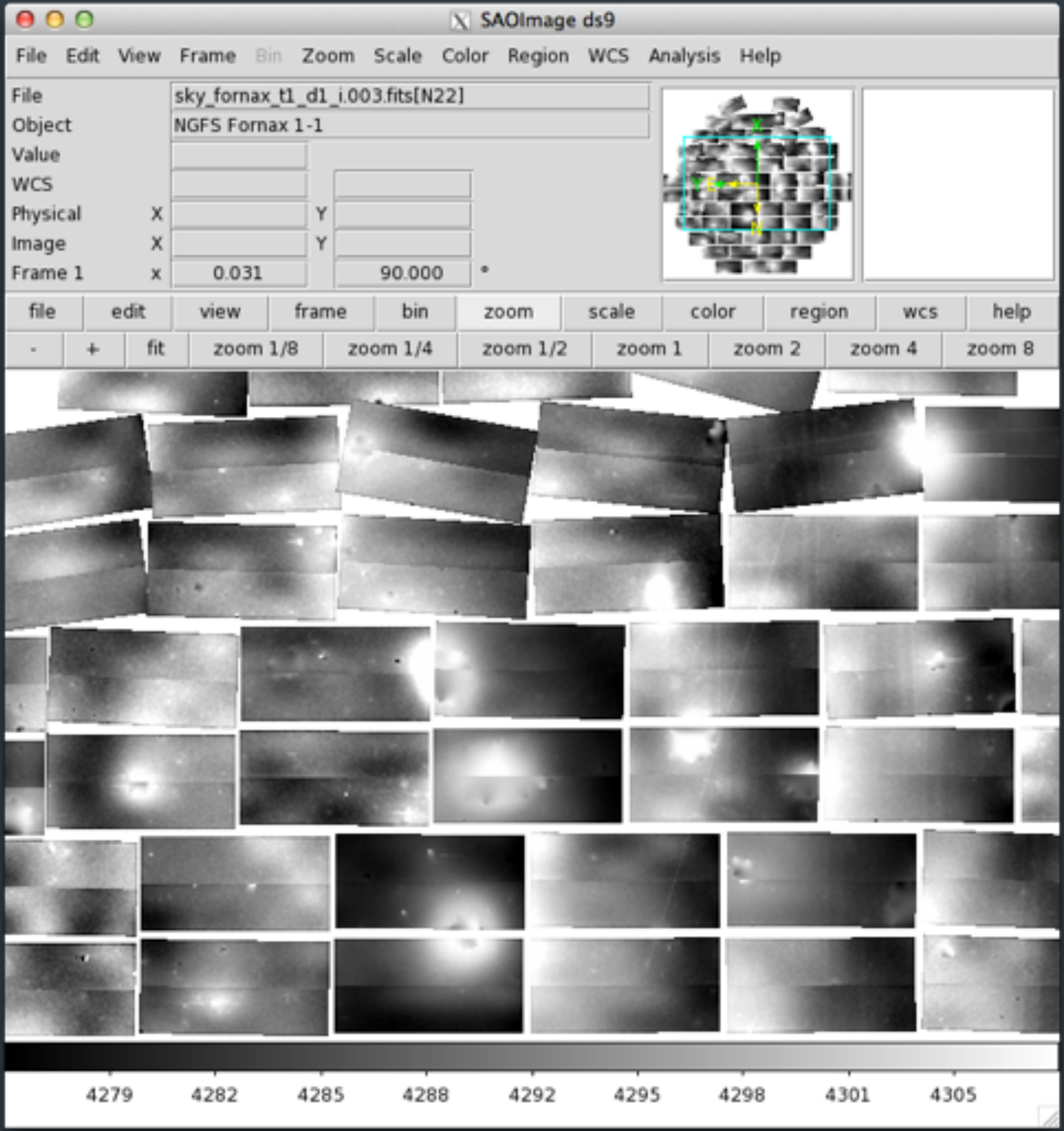
Sky subtraction: Splines



SAOImage ds9

File masked_median_fornax_t1_d1_i.003.fits[N23]
Object NGFS Fornax 1-1
Value
WCS
Physical X Y
Image X Y
Frame 1 x 0.031 90.000 °

File Edit View Frame Bin Zoom Scale Color Region WCS Analysis Help
file edit view frame bin zoom scale color region wcs help
- + fit zoom 1/8 zoom 1/4 zoom 1/2 zoom 1 zoom 2 zoom 4 zoom 8
4277 4284 4292 4299 4307 4314 4322 4329 4337



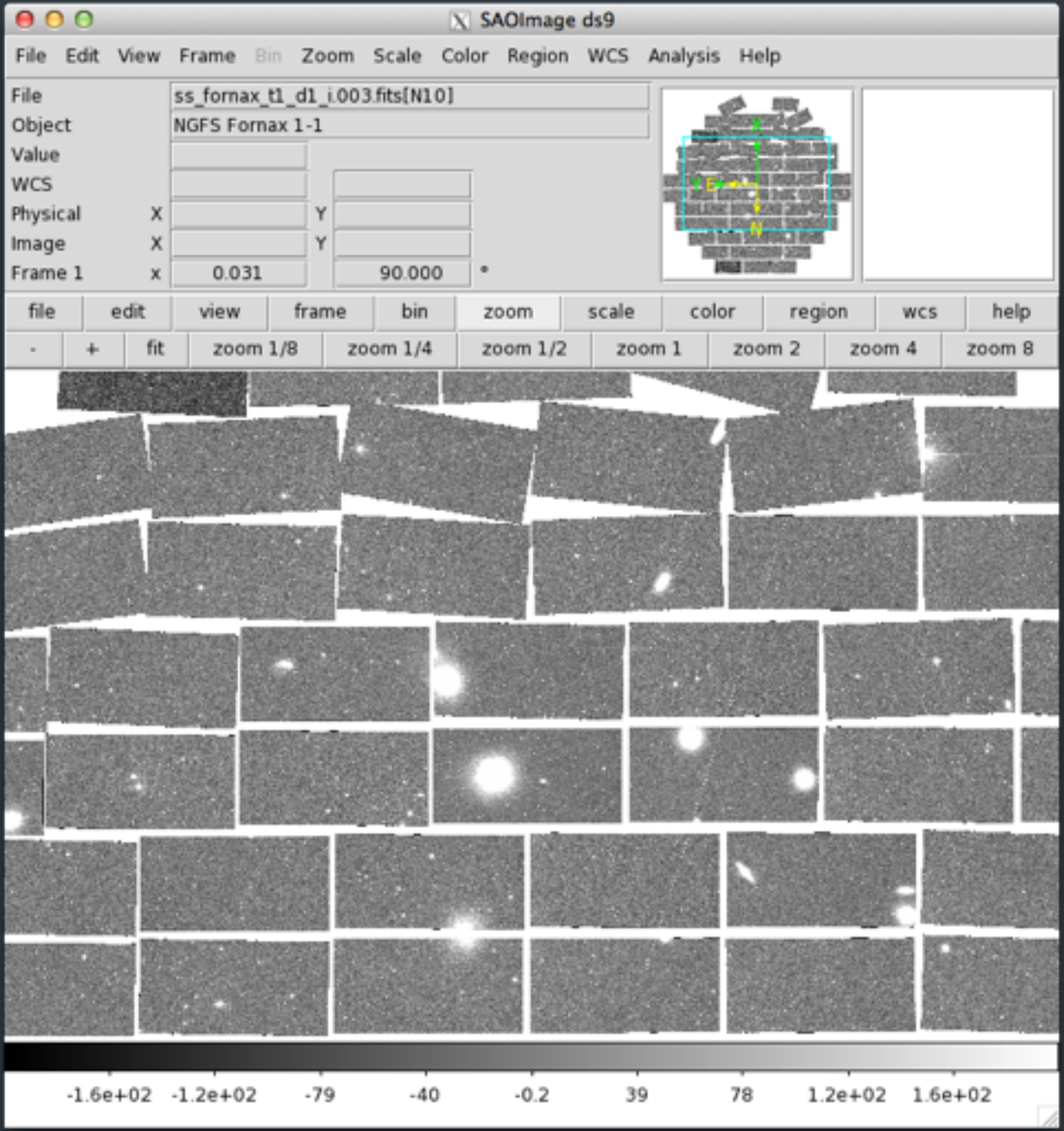
SAOImage ds9

File sky_fornax_t1_d1_i.003.fits[N22]
Object NGFS Fornax 1-1
Value
WCS
Physical X Y
Image X Y
Frame 1 x 0.031 90.000 °

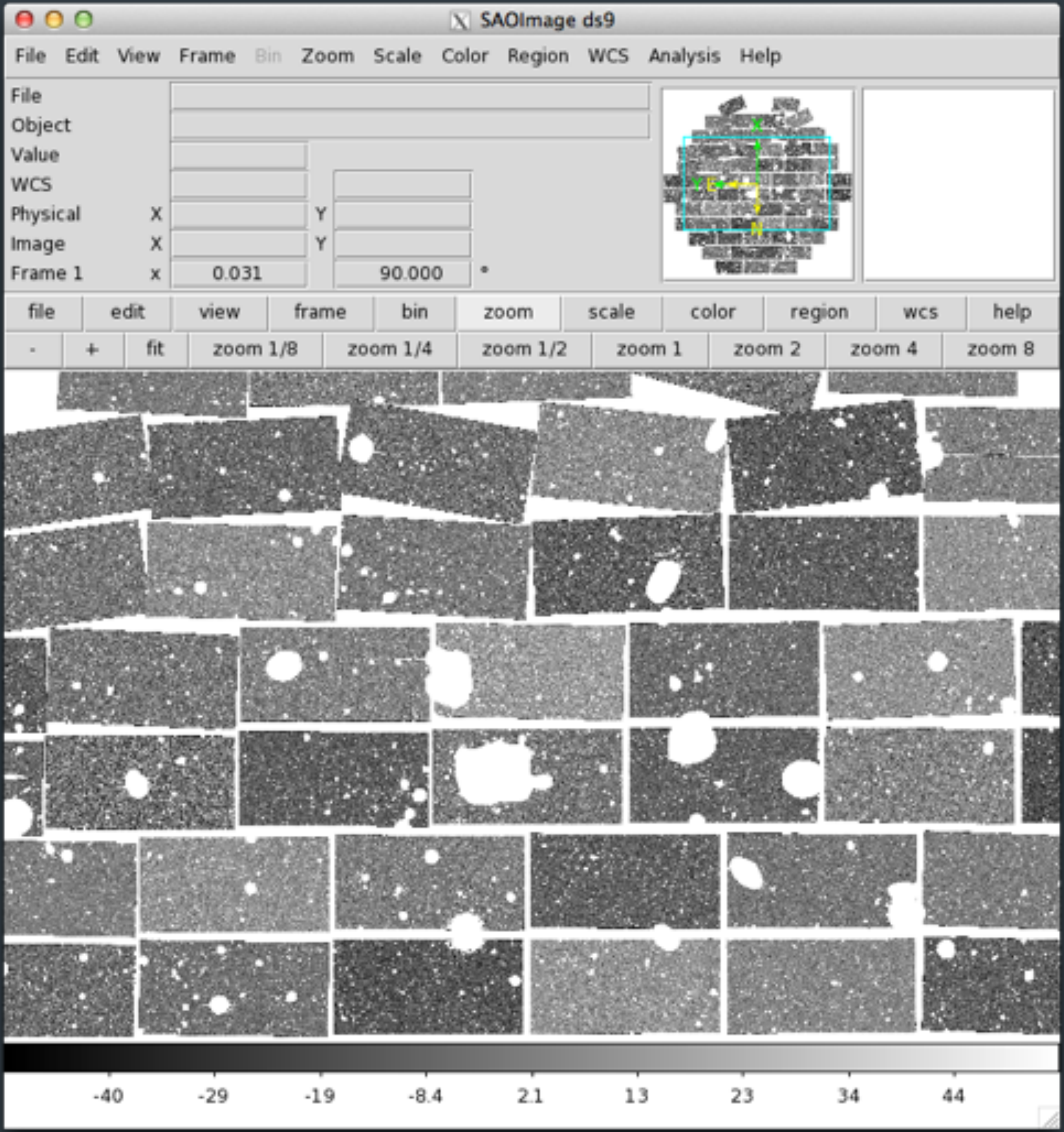
File Edit View Frame Bin Zoom Scale Color Region WCS Analysis Help
file edit view frame bin zoom scale color region wcs help
- + fit zoom 1/8 zoom 1/4 zoom 1/2 zoom 1 zoom 2 zoom 4 zoom 8
4279 4282 4285 4288 4292 4295 4298 4301 4305

Roberto Muñoz
Tools for Astronomical Big Data

Sky subtraction: Splines

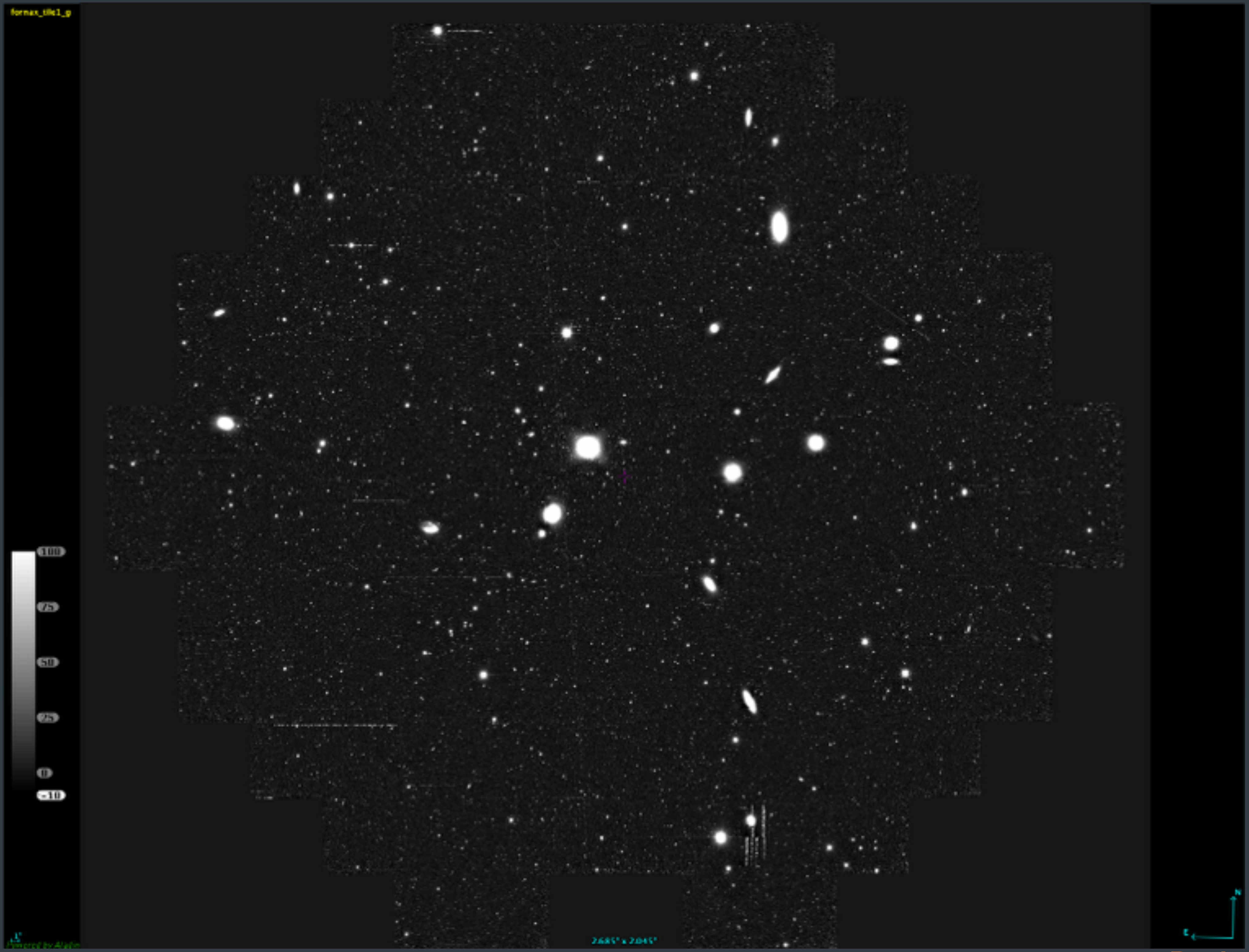

File: ss_fornax_t1_d1_i.003.fits[N10]
Object: NGFS Fornax 1-1
Frame 1 x 0.031 90.000 °

SAOImage ds9
File Edit View Frame Bin Zoom Scale Color Region WCS Analysis Help
Value
WCS
Physical X Y
Image X Y
file edit view frame bin zoom scale color region wcs help
- + fit zoom 1/8 zoom 1/4 zoom 1/2 zoom 1 zoom 2 zoom 4 zoom 8
-1.6e+02 -1.2e+02 -79 -40 -0.2 39 78 1.2e+02 1.6e+02

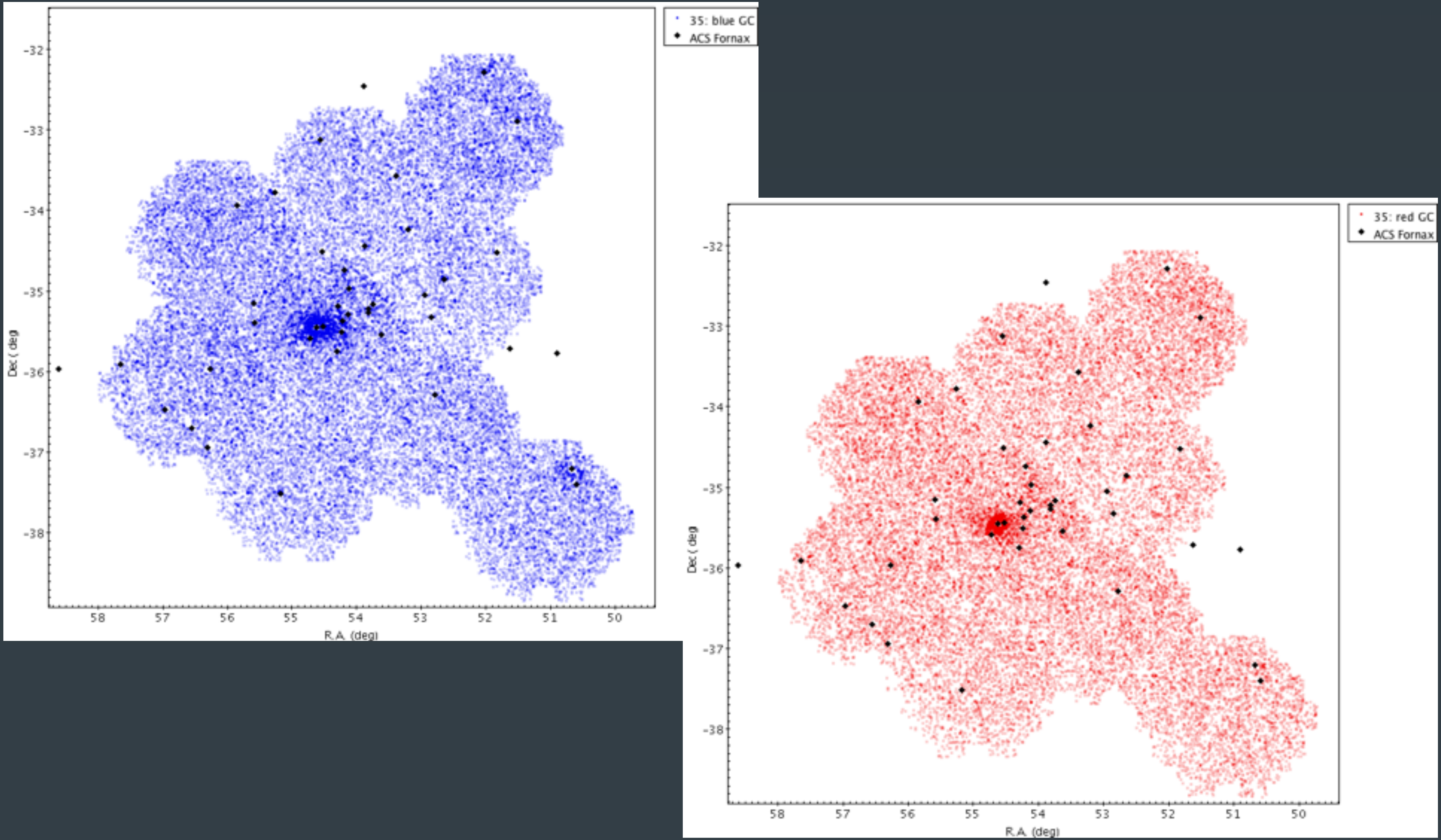

Frame 1 x 0.031 90.000 °

SAOImage ds9
File Edit View Frame Bin Zoom Scale Color Region WCS Analysis Help
File
Object
Value
WCS
Physical X Y
Image X Y
file edit view frame bin zoom scale color region wcs help
- + fit zoom 1/8 zoom 1/4 zoom 1/2 zoom 1 zoom 2 zoom 4 zoom 8
-40 -29 -19 -8.4 2.1 13 23 34 44

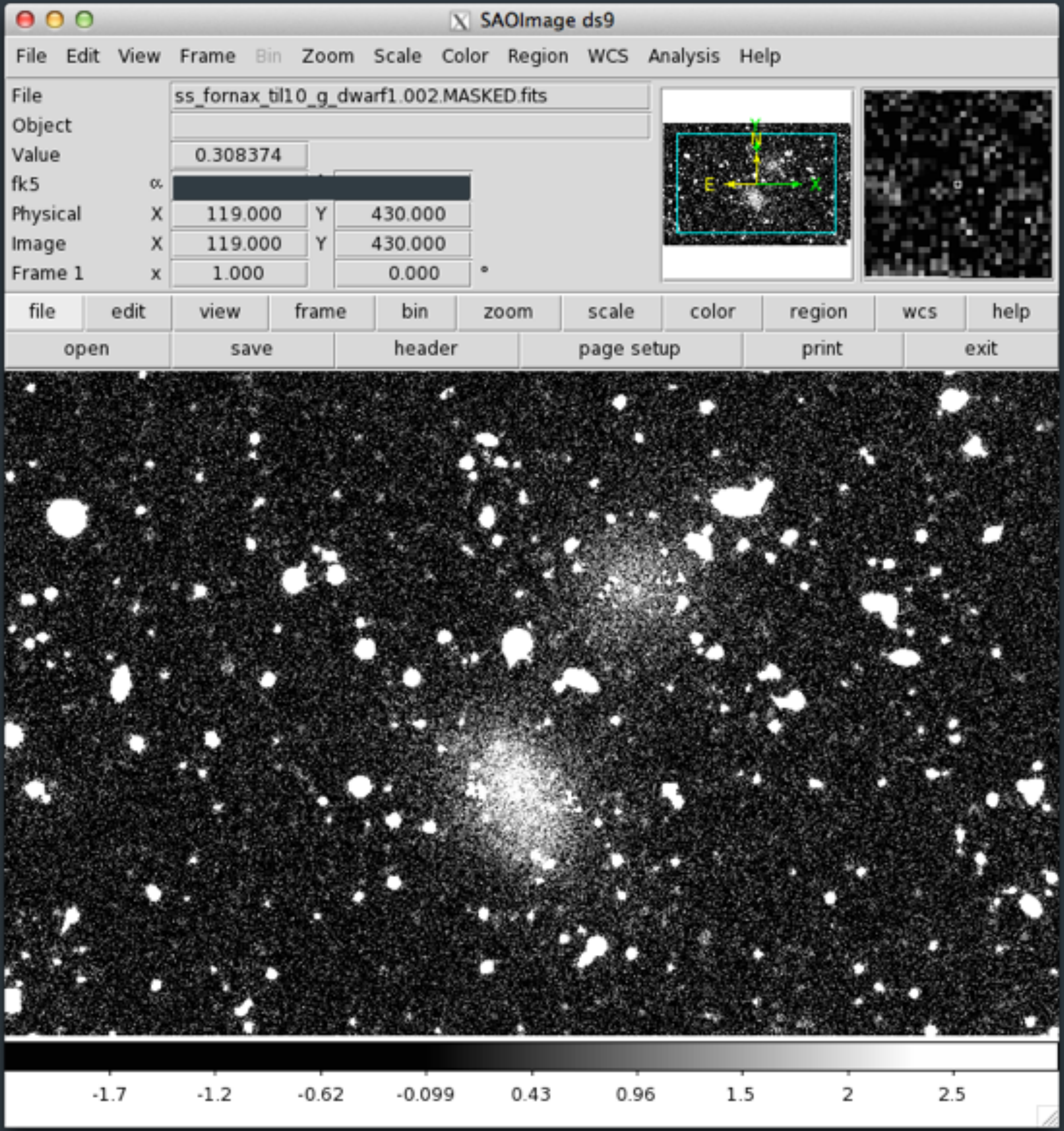
Stacks


fornax_tile1_g
100
75
50
25
0
-10
2.685° x 2.045°
N
E
Powered by Aladin

Fornax GCs


35: blue GC
ACS Fornax
Dec (deg
-32
-33
-34
-35
-36
-37
-38
58
57
56
55
54
53
52
51
50
R.A. (deg)
35: red GC
ACS Fornax
Dec (deg
-32
-33
-34
-35
-36
-37
-38
58
57
56
55
54
53
52
51
50
R.A. (deg)

Dwarf galaxies


SAOImage ds9
File Edit View Frame Bin Zoom Scale Color Region WCS Analysis Help
file edit view frame bin zoom scale color region wcs help
open save header page setup print exit
N
E
X
Y
-1.7
-1.2
-0.62
-0.099
0.43
0.96
1.5
2
2.5

File ss_fornax_til10_g_dwarf1.002.MASKED.fits

Object

Value 0.308374

fk5 α [redacted]

Physical X 119.000 Y 430.000

Image X 119.000 Y 430.000

Frame 1 x 1.000 0.000 °

Summary

- Setting up a large storage system in your office or institute to serve all your team 24 hours a day / 7days a week is possible.
- Plenty of disk space, high I/O and fast cpu processing for testing new algorithms and designing new methods. Different type of users in your team, some of them have lot of experience and others ones are still learning.
- Doing image processing and developing new methods take time, but the payback is good because you learn about your dataset and push it to the limits.
- Different projects and users require different solutions. We should support different approaches and look for the best solution depending on the needs and resources.