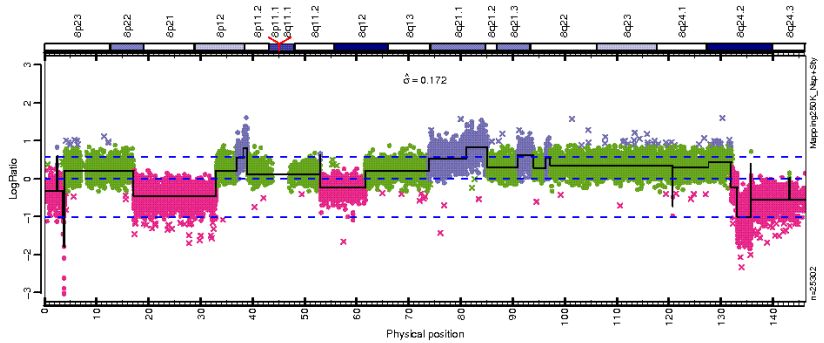
aroma.affymetrix - Processing very large Affymetrix data sets in bounded memory



Henrik Bengtsson

Postdoctoral researcher, Dept of Statistics, UC Berkeley.
(PhD Mathematical Statistics, MSc Computer Science)

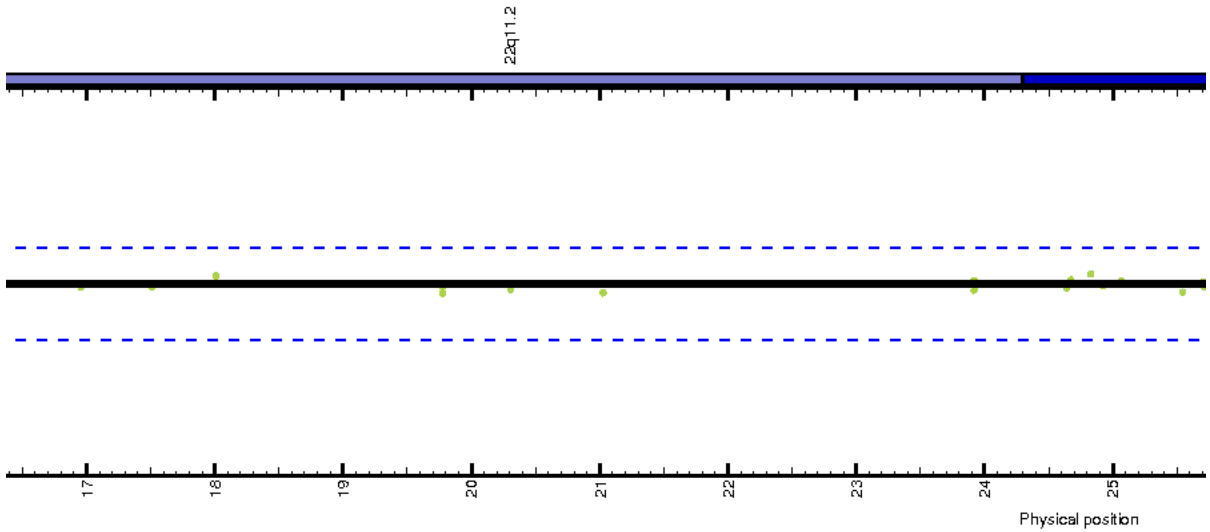
April 24, 2008

BioC2008, Lausanne
Session: Large Datasets
(slides will be available on the aroma.affymetrix webpage)

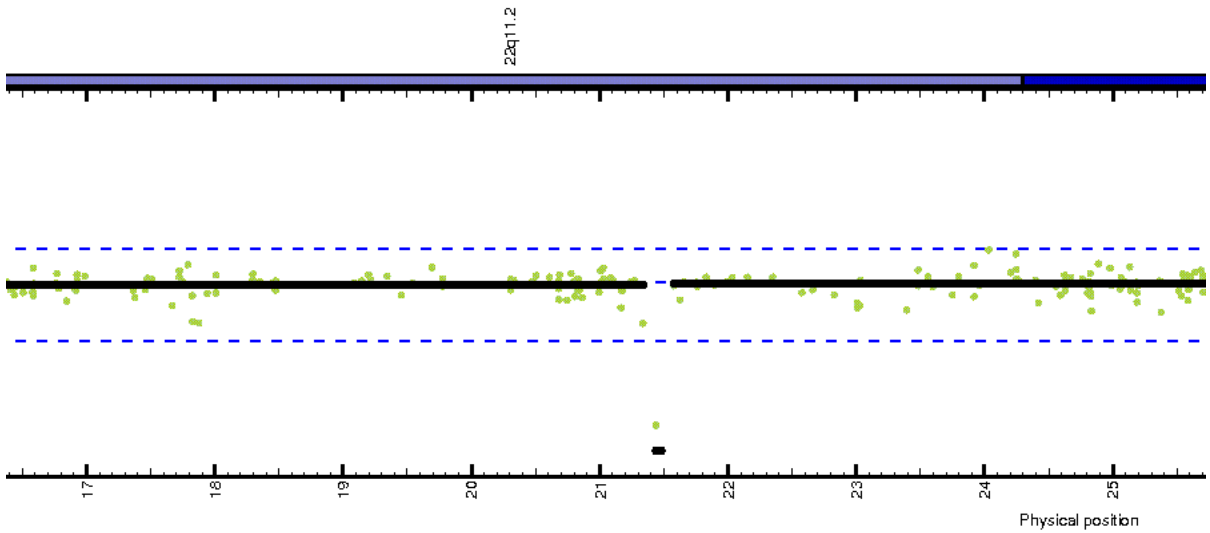
Impressive increase
in resolution

2003: 10,000 loci x1

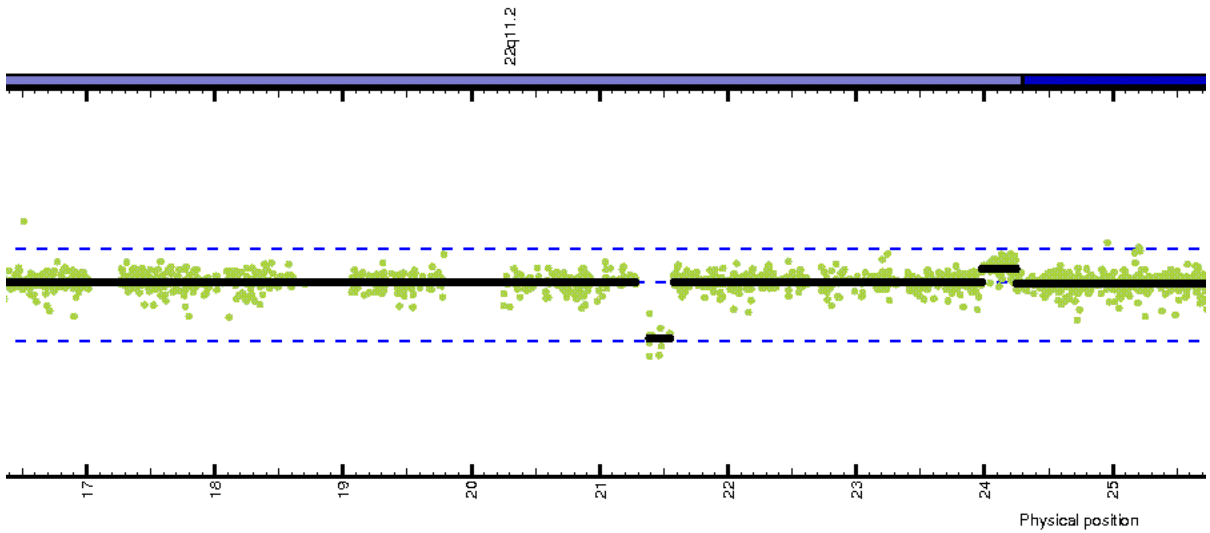
Example: 9Mb on Chr22



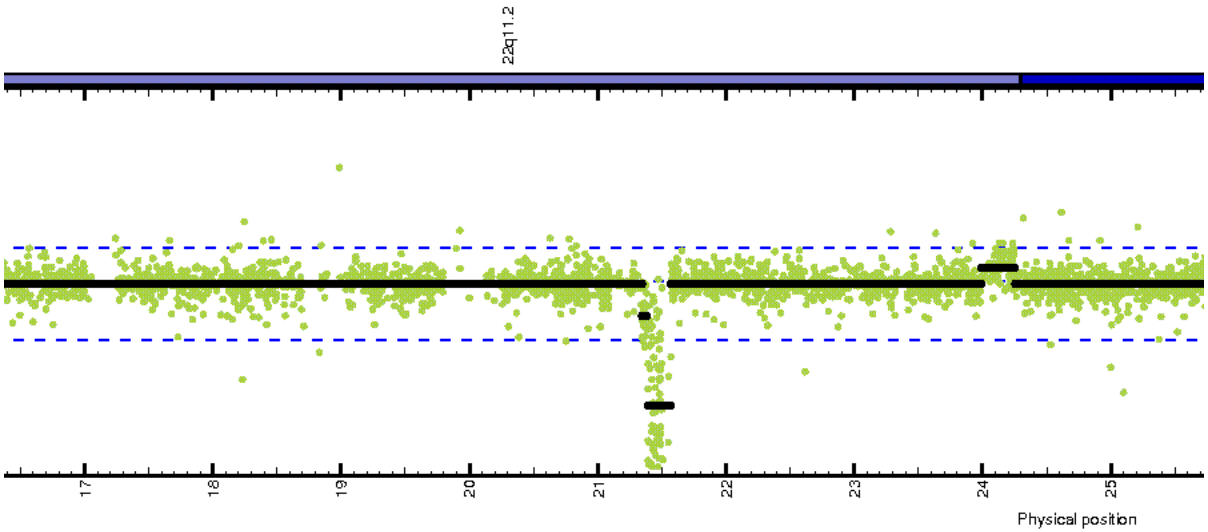
2004: 100,000 loci x10



2005: 500,000 loci x50



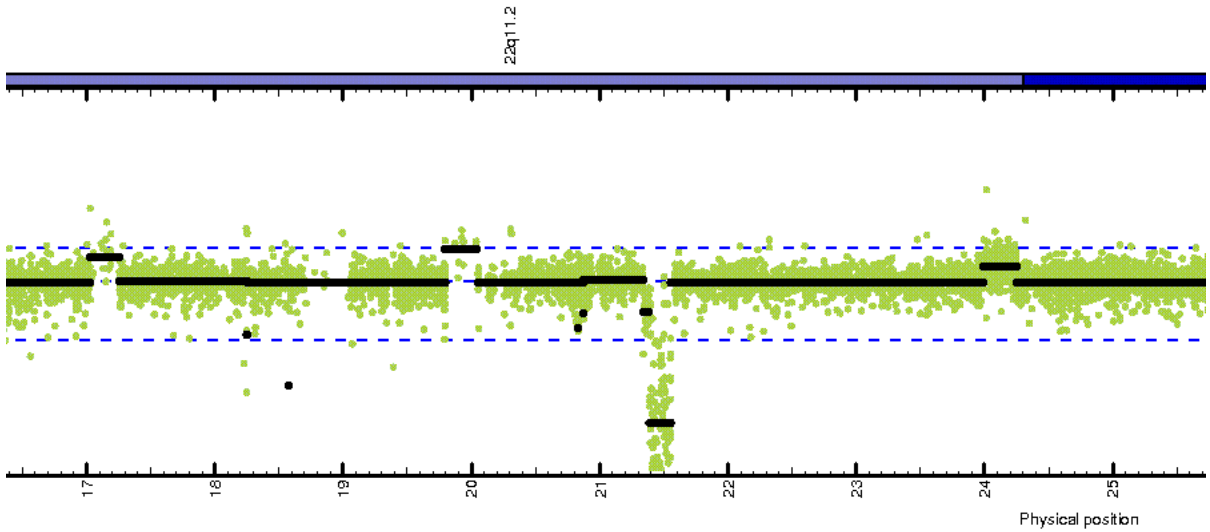
2006: 900,000 loci x90



TODAY

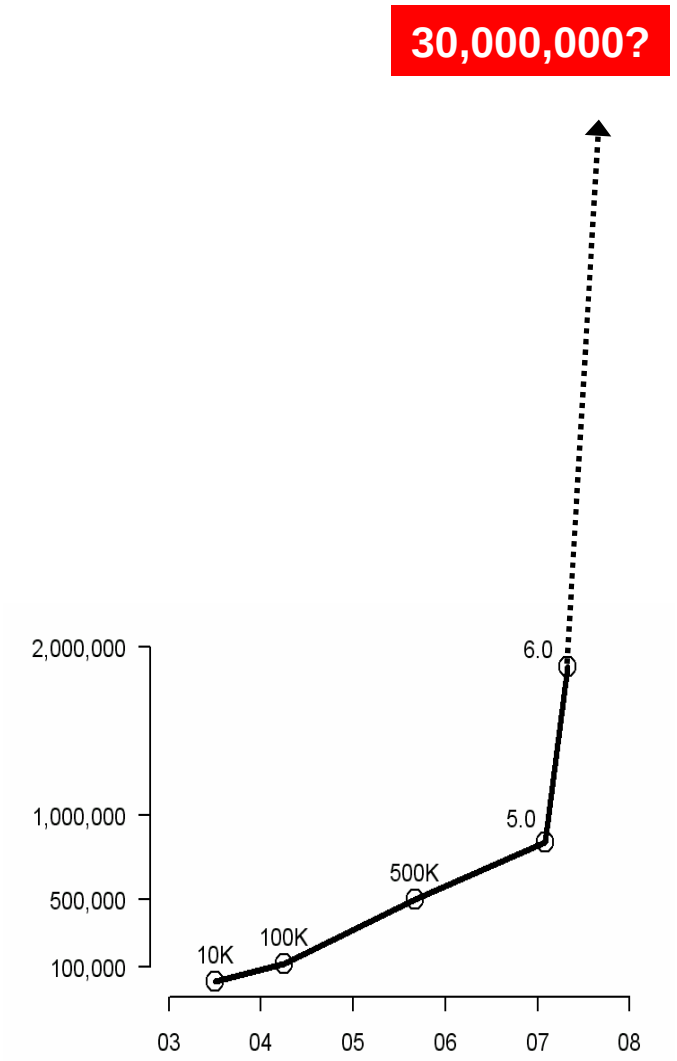
2007: 1,800,000 loci x180

Price: 4,700 loci for an espresso = a ~10K chip for two lattes
(Berkeley, CA, Spring 2008)



Affymetrix CEO, January 9, 2008:

2008(?): 30,000,000 loci x3000

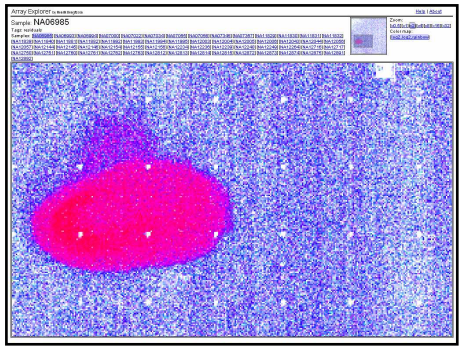


File sizes:

6.0:	70 MB/chip
5.0:	50 MB/chip
500K:	130 MB/chip set
100K:	50 MB/chip set
10K:	5 MB/chip

Overview of aroma.affymetrix

aroma.affymetrix processes unlimited number of arrays



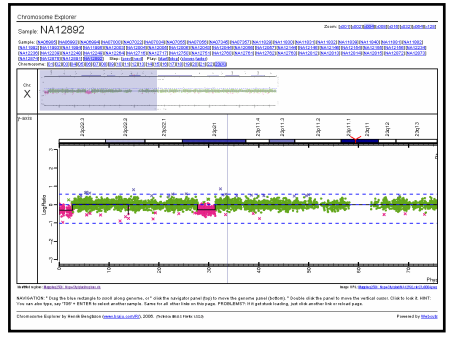
- Processes **unlimited number of arrays**:
  - Bounded memory algorithms.
  - Works toward file system.
  - Persistent memory: robust & picks up where last stopped.

- Memory requirements: **1.0-2.0GB RAM**.
  - Example: RMA on 4500 HG-U133A arrays uses ~500MB of RAM.
  - Example: CRMA on 300 SNP6.0 arrays uses ~1.5GB of RAM.
  - Example: FIRMA on 200 HuEx-1.0 arrays uses ~1.5GB of RAM.

- **Cross platform**: Linux/Unix, Windows, OSX.

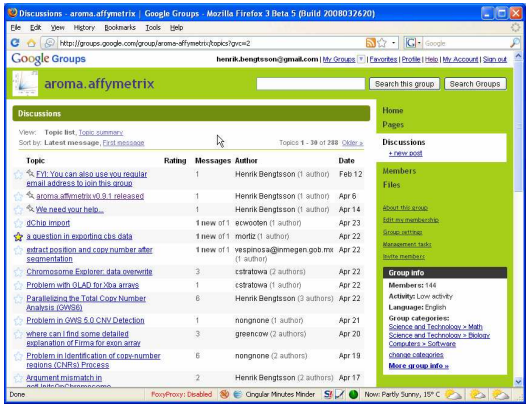
- Supports **most Affymetrix chip types**:
  - All chip types with a CDF (and some more).
  - Custom CDFs.

aroma.affymetrix "implements" several existing methods



- **Pre-processing**:
  - Background correction methods: RMA, gcRMA, ...
  - Allelic cross-talk calibration, quantile normalization, spatial normalization, ...
- **Probe-level summarization**:
  - multiplicative (dChip), affine, and log-additive (RMA) models.
- **Post-processing**:
  - PCR fragment-length normalization, GC-content normalization.
- **Quality assessment**:
  - RLE (*Relative Log Expressions)*, NUSE (Normalized Unscaled Standard Error)
  - Spatial plots: probe signals, PLM residuals, chip effects, CDF annotations, ...
- Paired & non-paired **copy-number analysis**:
  - All SNP & CN platforms. Multiple chip types.
  - CRMA (our methods for estimating raw CNs).
  - Segmentation method: CBS & GLAD. Easy to add more.
- **Alternative splicing**:
  - Finding Isoforms using RMA (FIRMA) (Purdom, Robinson, Simpson, Speed)

aroma.affymetrix is an open-source solution



- Community:
  - **200+ users** worldwide (approx 5-10 installation per day)
  - Active mailing list (Google Groups; 150+ message per month)
  - **Collaborative documentation** and vignettes (Google Groups).

- Development:
  - **2.5 years** since start:
  - Jan 2006-Oct 2006: Phase I: **Identifying API** (1-3 person project).
  - Oct 2006-Feb 2007: Phase II: **Maturing API** and testing (10-15 users & developers).
  - Feb 2007-Aug 2007: Phase III: **Extended real-world testing** (30-50 users & developers).
  - Aug 2007-...: Phase IV: **Public release** and heaps of CPU mileage (more "wild" use cases).

  - 3-5 active developers. One main maintainer / code coordinator.
  - Some external code snippet contributors.

  - Lots of **validation code** - catches existing and future bugs.
  - 1,000+ pages code / Rd pages.

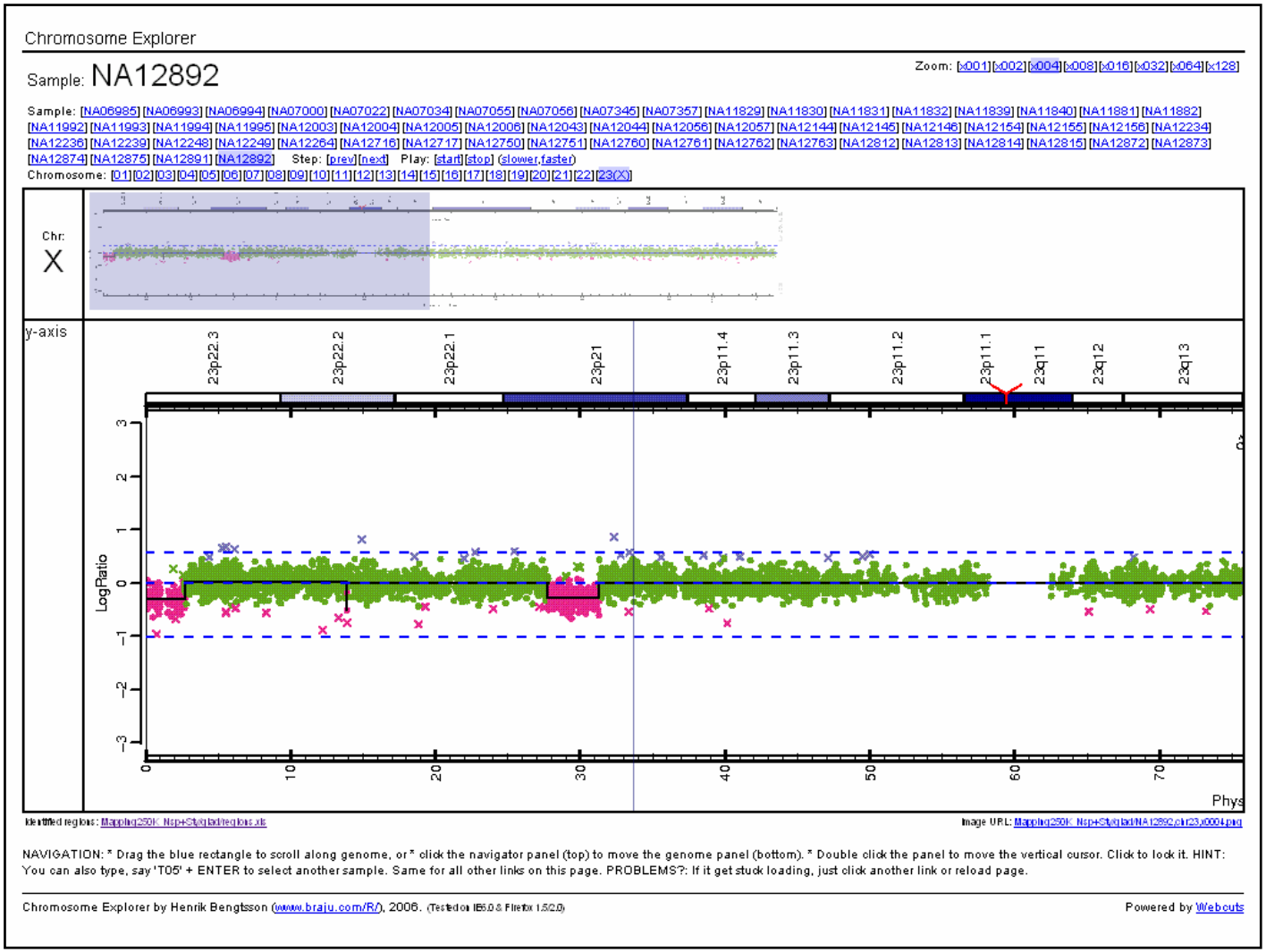
- Standards:
  - **Standard file formats**, e.g. reads/writes CEL (via **affxparser**/Fusion SDK).
  - Imports and exports to: APT, GTC, CNAG, CNAT, dChip, Bioconductor etc.
  - **Strict directory structures** and relative pathnames
    => portable scripts, robustness, more automated validations, easier to troubleshoot, simplified support.
  - Utilizes **existing packages**, e.g. preprocessCore, gcRMA, DNAcopy...

Walk-through example

Complete aroma.affymetrix script for copy-number analysis of 270 SNP6.0 samples

```
cs <- AffymetrixCelSet$byName("HapMap270",
                              chipType="GenomeWideSNP_6,Full")

acc <- AllelicCrosstalkCalibration(cs)
csC <- process(acc)

plm <- AvgCnPlm(csC, combineAlleles=TRUE)
fit(plm)

ces <- getChipEffectSet(plm)
fln <- FragmentLengthNormalization(ces)
cesN <- process(fln)

seg <- CbsModel(cesN)
fit(seg)

ce <- ChromosomeExplorer(seg)
process(ce)
```

Offline & online dynamic HTML reports Example: ChromosomeExplorer



Setup is as simple as placing the files in a strict & standardized directory structure

```
annotationData/
  chipTypes/
    GenomeWideSNP_6/
      GenomeWideSNP_6,Full.CDF
      GenomeWideSNP_6,Full.UGP
      GenomeWideSNP_6,Full.UFL

rawData/
  HapMap270,6.0,CEU,testSet/
     GenomeWideSNP_6/
       *.CEL
```

No (absolute) pathnames are used - maximizes portability

```
annotationData/
  chipTypes/
    GenomeWideSNP_6/
      GenomeWideSNP_6,Full.CDF  GenomeWideSNP_6,Full.UGP  ...

cdf <- AffymetrixCdfFile$byName("GenomeWideSNP_6", tags="Full")
print(cdf)

AffymetrixCdfFile:
Path: annotationData/chipTypes/GenomeWideSNP_6
Filename: GenomeWideSNP_6,Full.cdf
Filesize: 470.44MB
File format: v4 (binary; XDA)
Chip type: GenomeWideSNP_6,Full
Dimension: 2572x2680
Number of cells: 6892960
Number of units: 1881415
...
```

The file system is the memory
- data is loaded only when needed

```
cs <- AffymetrixCelSet$byName("HapMap270",
                chipType="GenomeWideSNP_6,Full")
print(cs)

AffymetrixCelSet:
Name: HapMap270
Tags:
Path: rawData/HapMap270/GenomeWideSNP_6
Chip type: GenomeWideSNP_6,Full
Number of arrays: 270
Names: NA06985, NA06991, ..., NA07019
Total file size: 17.7GB
RAM: 0.01MB
```

Normalized data is stored as CEL files - import to any software

```
acc <- AllelicCrosstalkCalibration(cs)
csC <- process(acc)
print(csC)
AffymetrixCelSet:
Name: HapMap270
Tags: ACC
Path: probeData/HapMap270,ACC/GenomeWideSNP_6
Chip type: GenomeWideSNP_6,Full
Number of arrays: 270
Names: NA06985, NA06991, ..., NA07019
Total file size: 17.7GB
RAM: 0.01MB

files <- getPathnames(csC)
print(files[1])
[1] "probeData/HapMap270,6.0,CEU,testSet,ACC,ra,
     -XY/GenomeWideSNP_6/NA06985.CEL"
```

Data sets (directories) are marked with unique tags

```
qn <- QuantileNormalization(csC)
csN <- process(qn)
print(csN)
```

```
AffymetrixCelSet:
Name: HapMap270
Tags: ACC,QN
Path: probeData/HapMap270,ACC,QN/GenomeWideSNP_6
Chip type: GenomeWideSNP_6,Full
Number of arrays: 270
Names: NA06985, NA06991, ..., NA07019
Total file size: 17.7GB
RAM: 0.01MB
```

Memoization

Memoization speed up computational expensive tasks by caching to file

Argument values -> md5 key -> cache lookup/update:

```
library(R.cache)

mySlowFcn <- function(foo, bar=2, verbose=TRUE) {
  key <- list(method="mySlowFcn", foo=foo, bar=bar)
  res <- loadCache(key)
  if (is.null(res)) {

    # Computational expensive algorithm
    res <- ...

    saveCache(key, res)
  }
  res
}
```

Extending aroma.affymetrix

Methods can be added by subclassing - CBS segmentation via a simple wrapper

```
setConstructorS3("CbsModel", function(cesTuple=NULL, ...) {
  library("DNAcopy")
  extend(CopyNumberSegmentationModel(cesTuple=cesTuple, ...), "CbsModel")
})

setMethodS3("fitOne", "CbsModel", function(this, data, chr, ...) {
  nbrOfUnits <- nrow(data)
  data <- DNAcopy::CNA(genomdat=data[,"M"], maploc=data[,"x"],
                       chrom=rep(chr, nbrOfUnits))
  DNAcopy::segment(data, ...)
})


setMethodS3("extractCopyNumberRegions", "DNAcopy", function(fit, ...) {
  with(fit$output, {
    CopyNumberRegions(chromosome=chrom,
    start=loc.start, stop=loc.end, mean=seg.mean, count=num.mark)
  })
})
```

Future development

aroma.affymetrix is preparing for local-network parallelization

Given:

1. Shared persistent memory, i.e. file system.
2. Processing of arrays / units / probes in chunks.

Today:

1. Manually launch multiple hosts.
2. Run single-array methods (e.g. some normalization, segmentation methods) by manual/tedious coordination.

Tomorrow:

1. Run identical analysis script on multiple hosts.

Requires:

1. **A file-locking mechanism (HELP NEEDED!).**
2. **Randomized processing (of arrays / units / probes).**

Acknowledgments

- *UC Berkeley:*
- **James Bullard**
- **Kasper Hansen**
- **Elizabeth Purdom**
- Terry Speed

- *WEHI, Melbourne:*
- **Mark Robinson**
- **Ken Simpson**

- *John Hopkins, Baltimore:*
- Benilton Carvalho
- Rafael Irizarry

- *ISREC, Lausanne:*
- Pratyaksha “Asa” Wirapati

Affymetrix, California:
Ben Bolstad
Simon Cawley
Steve Chervitz
Harley Gorrell
Earl Hubbell
Luis Jevons
Chuck Sugnet
Jim Veitch
Alan Williams
...