Image Bioinformatics

Image semantics in life sciences research

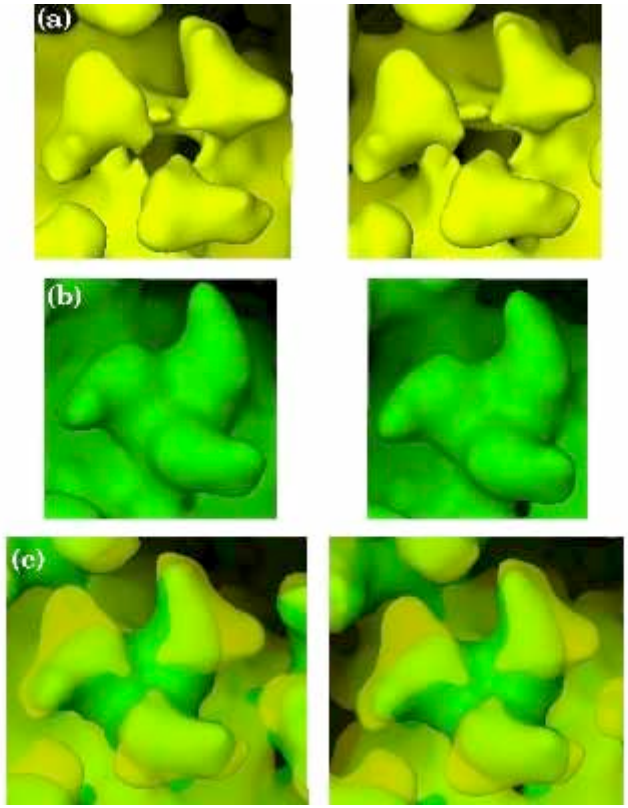
Graham Klyne
Image Bioinformatics Research Group of the Oxford e-Science Centre
Department of Zoology
University of Oxford, UK

Introduction
Image Bioinformatics Research Group

- We are: David Shotton, Chris Catton, Graham Klyne, Liz Mellings, based in the Zoology Department at Oxford
- Backgrounds in: cell biology, microscopy, animal behaviour, video, database design, ontology, Internet and Web standards, Semantic Web, data curation
- Part of the Oxford e-Science Centre
- Drawing on expertise and standards in biology, computing and ontologies, applied to life-sciences research

Technical Goals

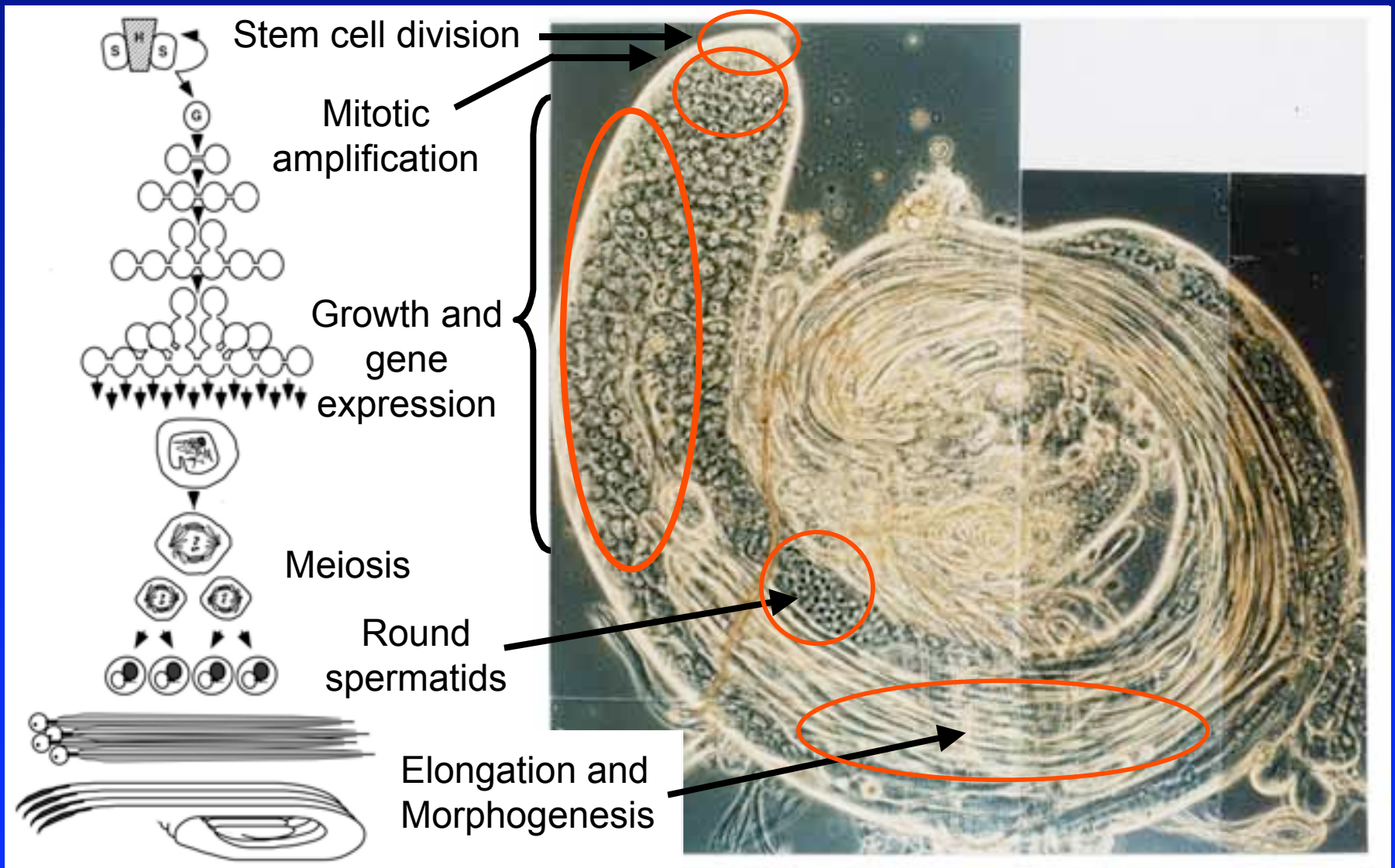
- Storage technology is ***not*** a goal for us
- Assembly of systems to capture and publish research images and metadata with associated high-level descriptions
- Preserving the association between raw data and high-level descriptions
- Provide access to data in terms of research domain concepts
- Combine research image data with other online resources (gene databases, literature databases, *etc.*)
- Web-style interoperability and evolveability



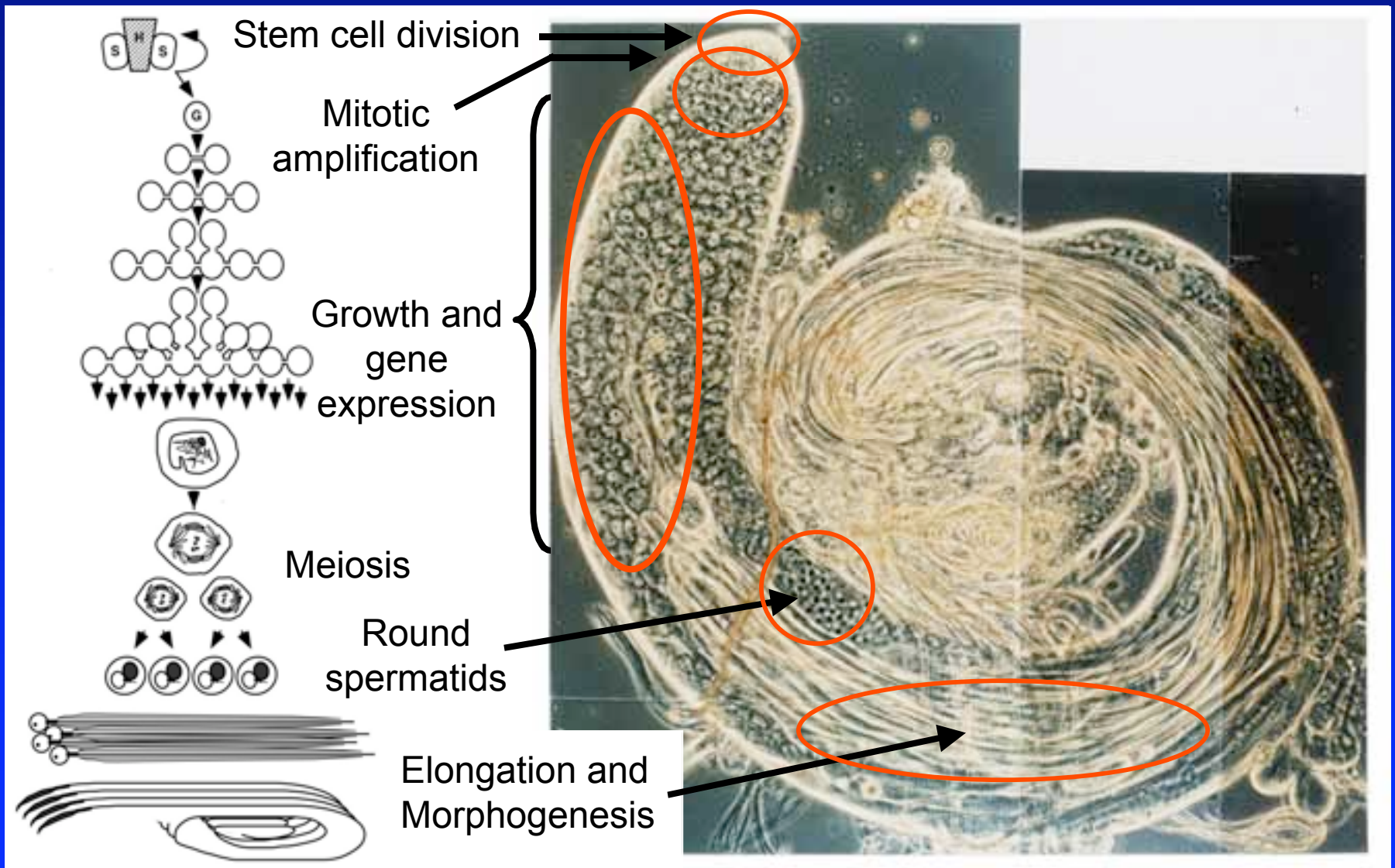
Drosophila Testis Gene Expression Database (DTGED) Project

- Research the function of genes whose expression is dependent on specific (*aly*-class) proteins
  - PIs: Dr Helen White-Cooper and Dr David Shotton
  - We are working closely with the DTGED research team based in the Zoology Department, University of Oxford
- Genes code for the production of complex chemicals (enzymes, proteins, etc) used in biological processes
  - But the *expression* of any gene is dependent on the cell environment, including the presence of other gene products
  - Observable biological consequences (phenotypes) may result from subtle interactions between many gene products and other factors
  - This project aims to document such interactions in *drosophila* (fruit fly) spermatogenesis

Adult *Drosophila* testis



Adult *Drosophila* testis



Images of Expression Patterns

- To an expert observer these images clearly show gene expression at different stages of spermatogenesis
- Each image corresponds to a different combination of gene and a strain of *drosophila*
- These *in situ* hybridization images are the "end game" - the final stage of a non-trivial process of screening and preparation
- Reproducibility and interpretation requires that the preparatory steps are recorded along with the images

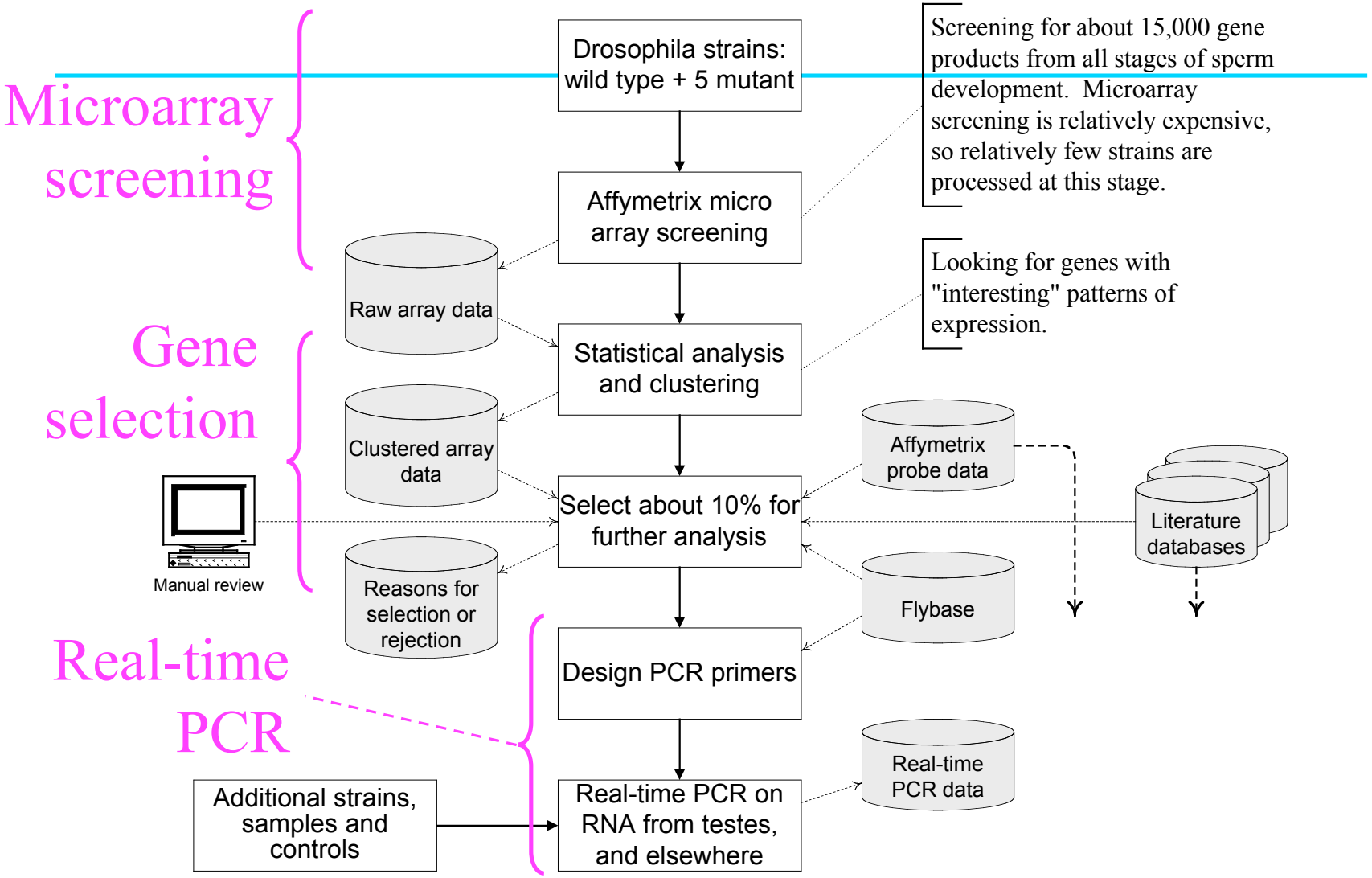
CG2247 wt

CG2247 topi

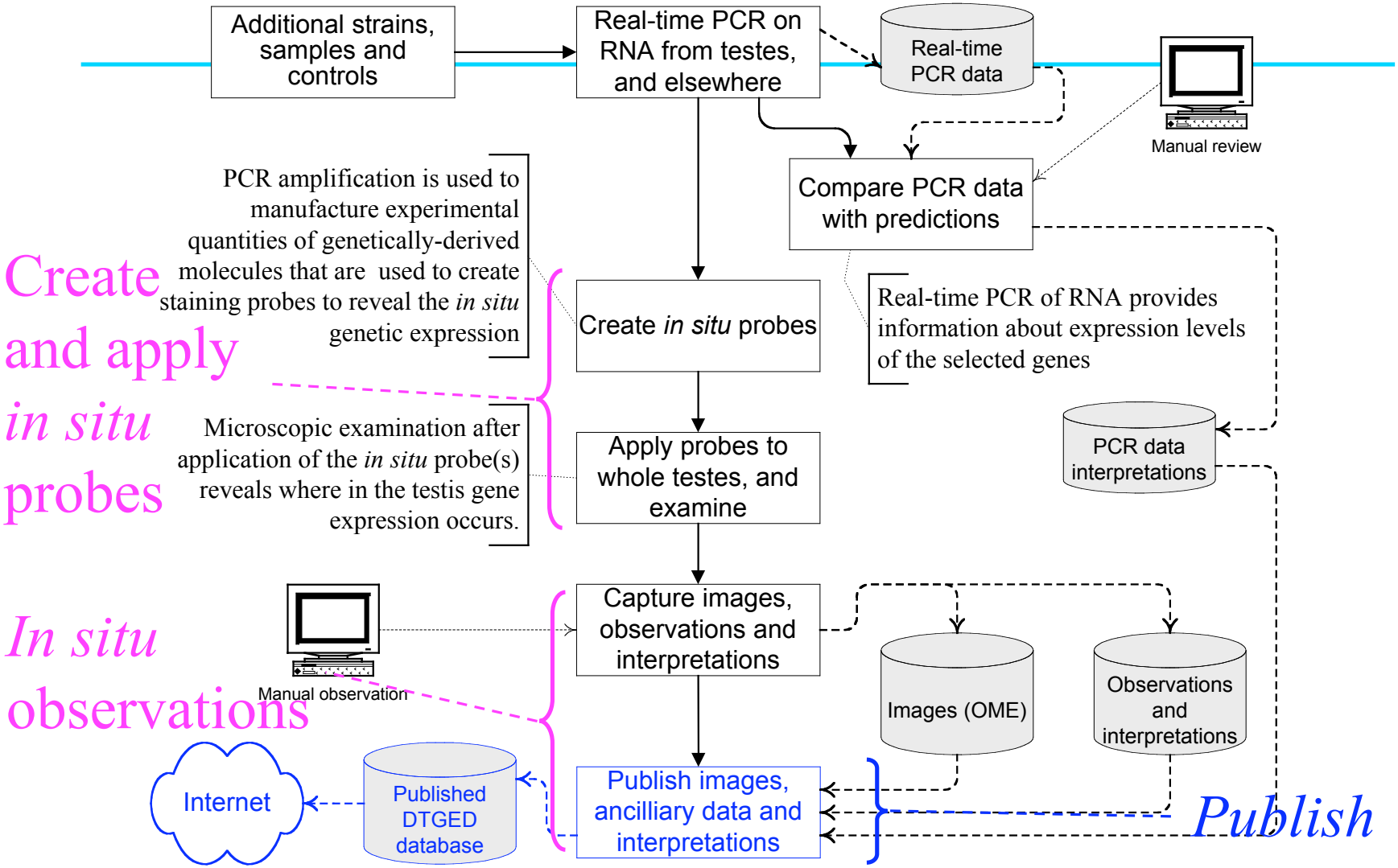
CG12907 aly

CG12907 topi

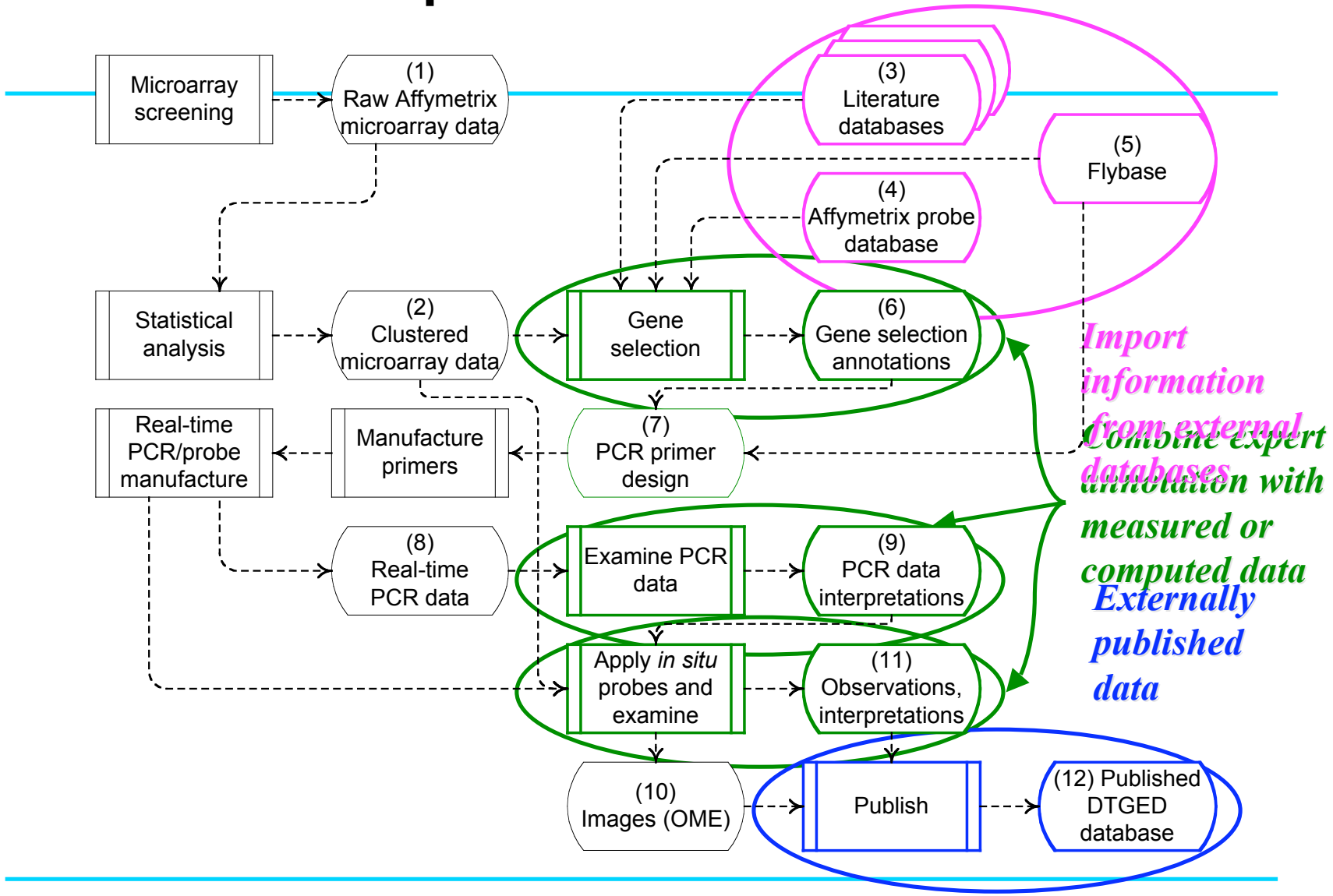
So how is it done? (1)



So how is it done? (2)



DTGED Experimental Data Flows



DTGED Technologies Used

- Minimalist approach to development; working with available web-based tools, etc.
- Protégé+Racer (DL reasoner) for design and testing of ontologies for experimental data
  - Note that expert annotations are open-ended
- BioImage for Ontology-directed capture and staging of annotations and observations
  - Extends original purpose of BioImage
- Open Microscopy Environment (OME) for capture and staging of images and image metadata
- Haskell for "conditioning" Excel spreadsheet data and combining it with other data sources
- BioImage for publication of images and metadata

Experiences with OME

- Our focus:
  - ***Manual*** annotation of images rather than high-volume automated processing
  - Integrate image annotations with other data used in experimental workflows
- Our experience (as of 3-4 months ago):
  - OME can be a difficult system to get into
  - Complex, monolithic system
  - Documentation seems to assume audience already "on message"
  - Limited and awkward in area of user annotation
  - Steep learning curve for even simple project-specific annotations
  - Problems with export of annotations
  - Data model is not easy to understand (2 forms?)

Desiderata for the Future

- OME/XML is the key component
  - the rest is "mere engineering"
- Clarify what the OME core needs to do
  - move anything that doesn't need to be there
  - minimize the core storage engine that needs to scale
  - provide other functions as "satellite" applications
- Where possible, design for scaling at the edge
  - No need to handle UI functions in the core database
- Make it easy for users to enlarge the system
  - Note: I would not choose to use Perl or Java for project-specific data handling (Python, Haskell, Ruby, or even Excel, are quicker for project-specific applications)
- "Small pieces, loosely coupled" (software *and* data mode)

Making Interoperability Easy

- Well-defined data format and semantics (OME/XML)
- Provide simple interfaces to access/update stored data
  - e.g. simple HTTP GET/POST
- Provide stand-alone tools and libraries to manipulate OME/XML data
- Thoroughly document the interfaces and data formats, including assumptions and limitations
- Plan to interwork with modern lightweight frameworks for Web UI work
  - e.g. Ruby/Rails, Python/Turbogears
    - Ajax for responsive possibilities
- Enable users to add functionality using any tools

Open Microscopy Web ?

- Enable researchers to review and find new results using existing published data - but first it must be published!
- The Web does not have a centre
- OME providing nodes in a web of research image data
- Separate publication from quality/review process
  - Use the web for raw publication
  - Anyone can publish data on the web, regardless of quality
  - Others can review and index data that meets appropriate criteria (technical standards, trust in researchers, peer review, etc.)
- Develop tools to combine disparate research information
  - Syntax translation to common form for exchange
  - Use ontologies and ontological engineering techniques to find correspondences