Proteomics and genomics standards - A biologist’s perspective

Jef D. Boeke

"Didn't I read in the paper the other day where they'd finally found out what the basic secret of life was?" "I missed that," I murmured. "I saw that," said Sandra. "About two days ago." "That's right," said the bartender. "What is the secret of life?" I asked. "I forget," said Sandra.

"Protein," the bartender declared. "They found out something about protein." "Yeah," said Sandra, "that's it."

-Kurt Vonnegut, Cat’s Cradle

For a given yeast protein/gene, what is the first stop for a biologist for deciphering protein function? (a highly personal view as a consumer of gene/protein information)

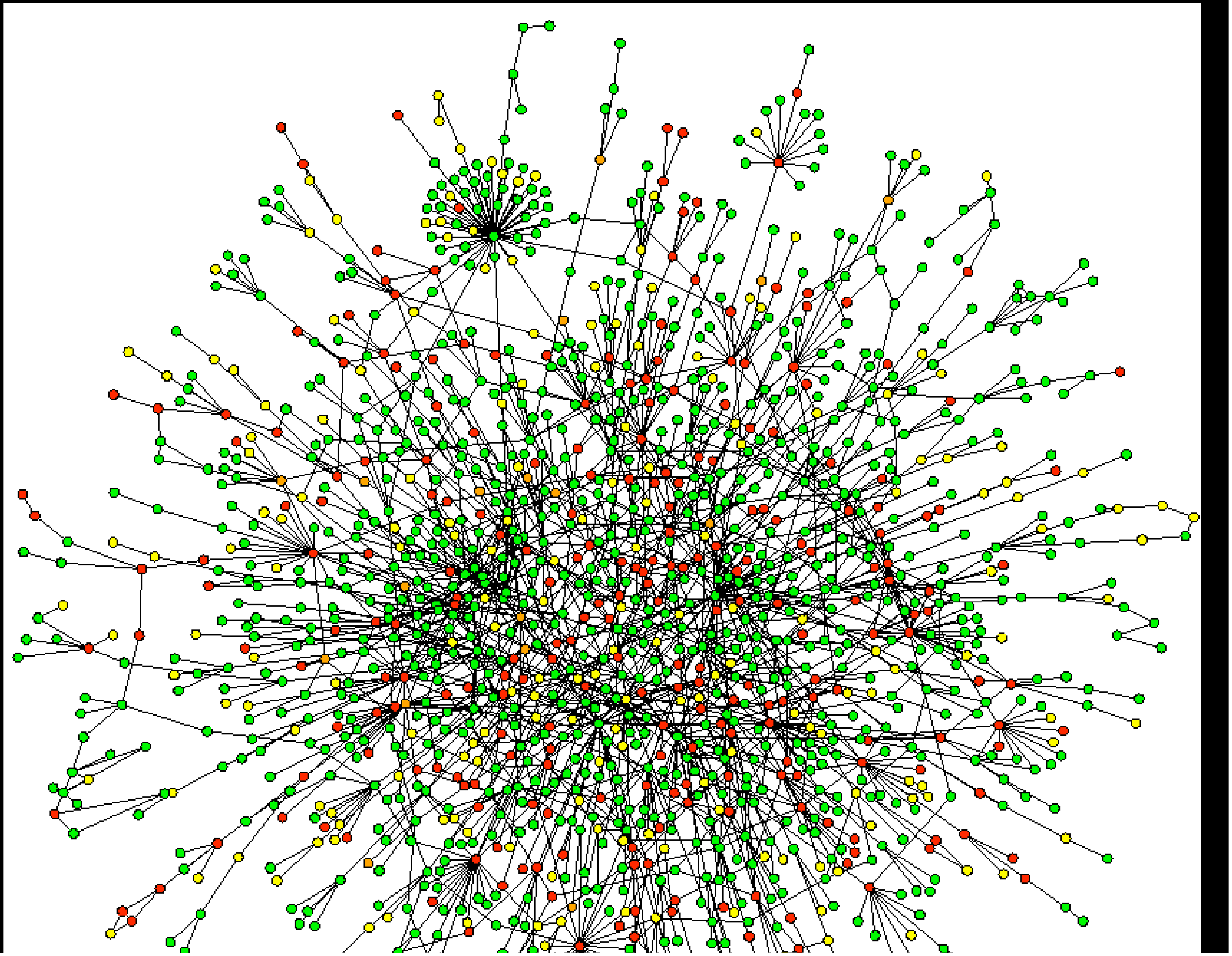
- YPD - yeast protein database - nonpublic, simple, one-stop; human curation v. valuable but nonscalable… Has extremely valuable human-curated “sound bites” particularly regarding the phenotypes of mutants, something that is not available elsewhere - answers the question “what happens when I take this protein away?
- SGD - Saccharomyces genome database - public, complex, focus on genome not proteome, does not attempt to seriously integrate large scale datasets
- Pubmed - Clearly essential, but not as useful as “sound bites” provided by YPD - too much information/unsystematic

Databases of interactions among proteins come second…

- GRID - focused on simple interaction lists, clean simple format

- BIND - most impressive attempt to systematically integrate large and small datasets, innovative icons. But steep learning curve limits usefulness… Automated parsing/curation of literature data can be highly misleading

Networks, network integration



Networks, network integration, viewpoint of a producer of high-throughput data

Well-defined standards are essential for deducing meaningful biological networks and pathways from high throughput datasets

High throughput datasets incorporating as many canned standards borrowed from other databases have a higher likelihood of being integrated and linked to by other databases

This will increase the efficacy with which “network integrators” can do their work

Evaluating high throughput data

- Well-defined data quality metrics are critical - noise intrinsic to large datasets is well known
- Such metrics must be systematic and rankable to be useful
- Best datasets have well-defined validation controls
- Data should be provided at different levels of abstraction for different levels of use
- High level
- Medium level
- Raw data level
- Data metrics help solve common problem for those producing high throughput data: When to release data? Nobody wants to release poor quality data, but all want to make data available as quickly as possible.

The Miame experience

- Systematic definition of the Minimum information about a microarray experiment
- Allows replication of experiments, in principle
- Controlled vocabularies can be used to enforce standard data types (e.g. organism name)
- Embraced by journals
- Repositories - e.g. GEO, Arrayexpress provide long-term data warehousing and user-friendly interface

- Reality is that data in repositories is incomplete and can be unsystematic
- Data sharing among repositories is not working well, if at all
- Problems are more complex than warehousing sequence information due to multidimensionality
- Very much a work in progress

Our high throughput project to determine protein function: genome SLAM --- deducing protein function from knockout mutant behavior

- Major collaboration with Forrest Spencer, Joel Bader and Rafael Irizarry
- Goal: database of ~25,000,000 possible genetic interactions determined by “SLAM” synthetic lethality analyzed by microarray
- A picture of genome redundancy and in combination with protein interaction maps, a “wiring diagram” of the cell
- New insights into “quantitative traits”
- Database of candidate gene interactions underlying human disease

Synthetic lethality: what is it and does it tell us?

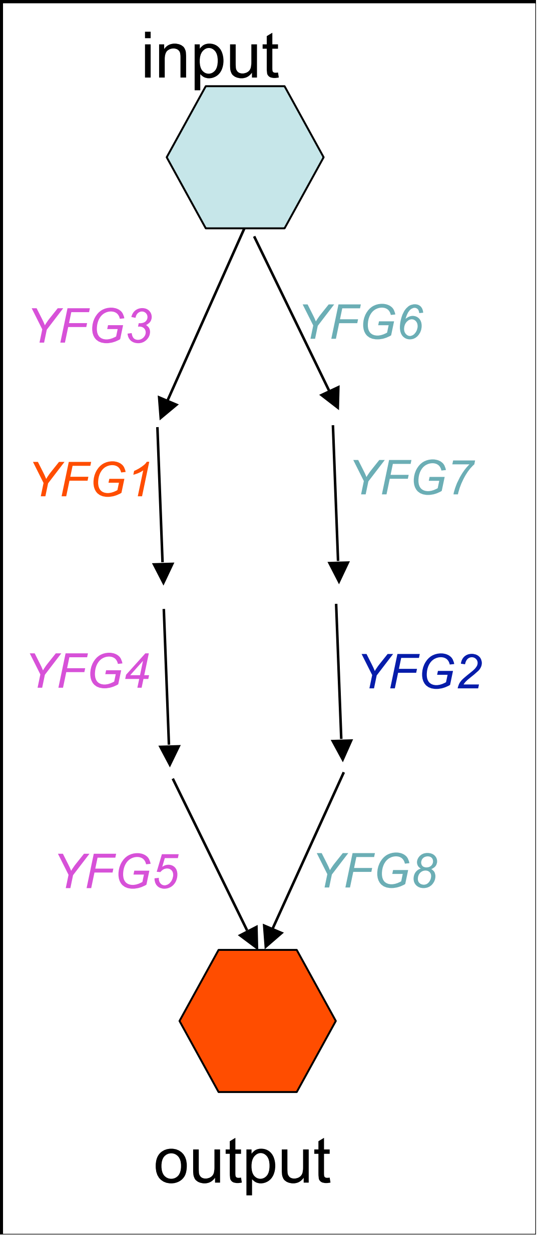
yfg1 mutant – viable
yfg2 mutant – viable
yfg1 yfg2 double mutant - inviable

If the nature of the *yfg* mutants is unknown, many possible interpretations…

BUT, if they are both null alleles, simplest interpretation is they are in redundant, parallel, or branched pathways

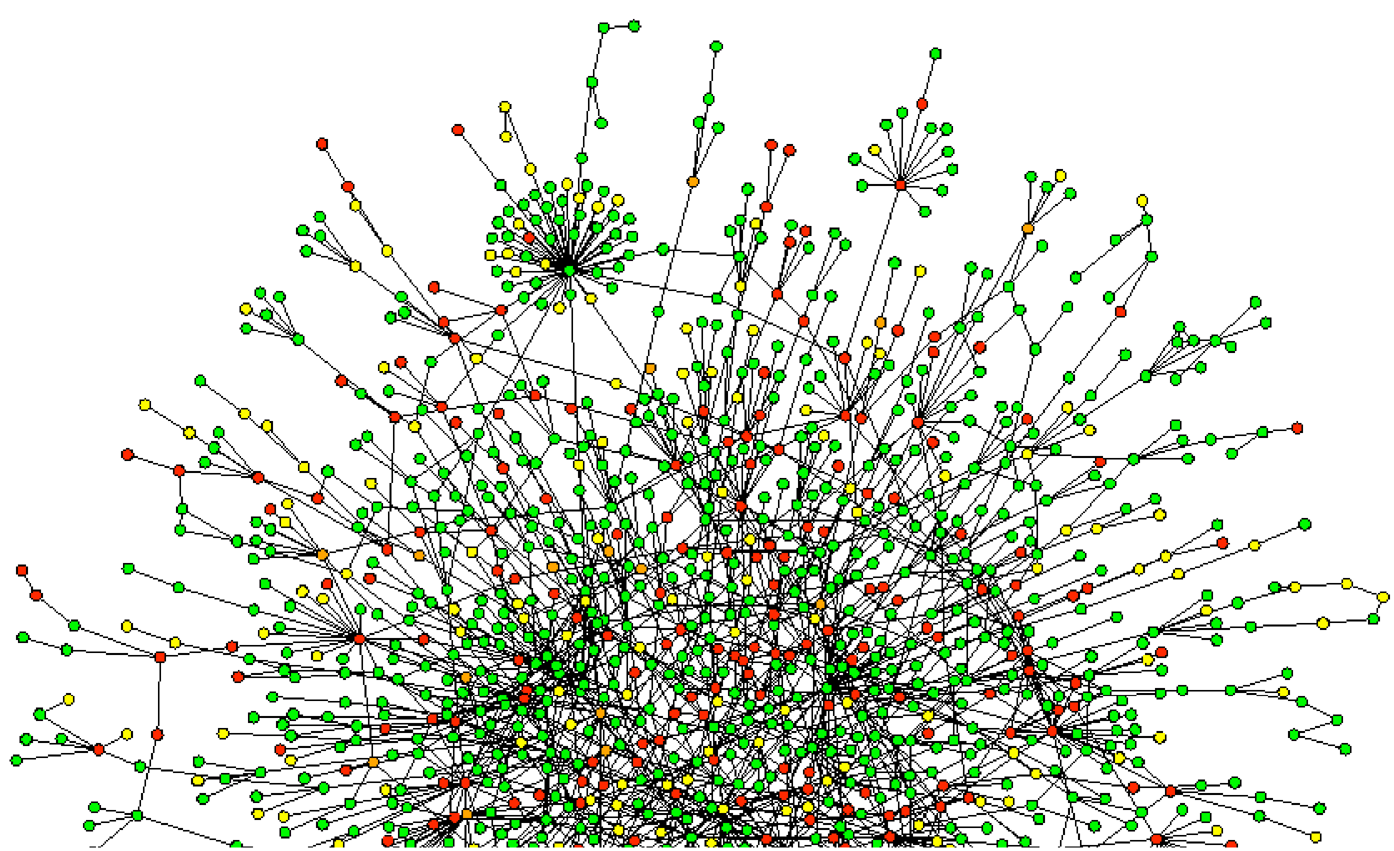
Thus, the patterns of lethality will help us deduce pathway architecture, especially in conjunction with protein interaction data

"Congruence Score" puts proteins into pathways

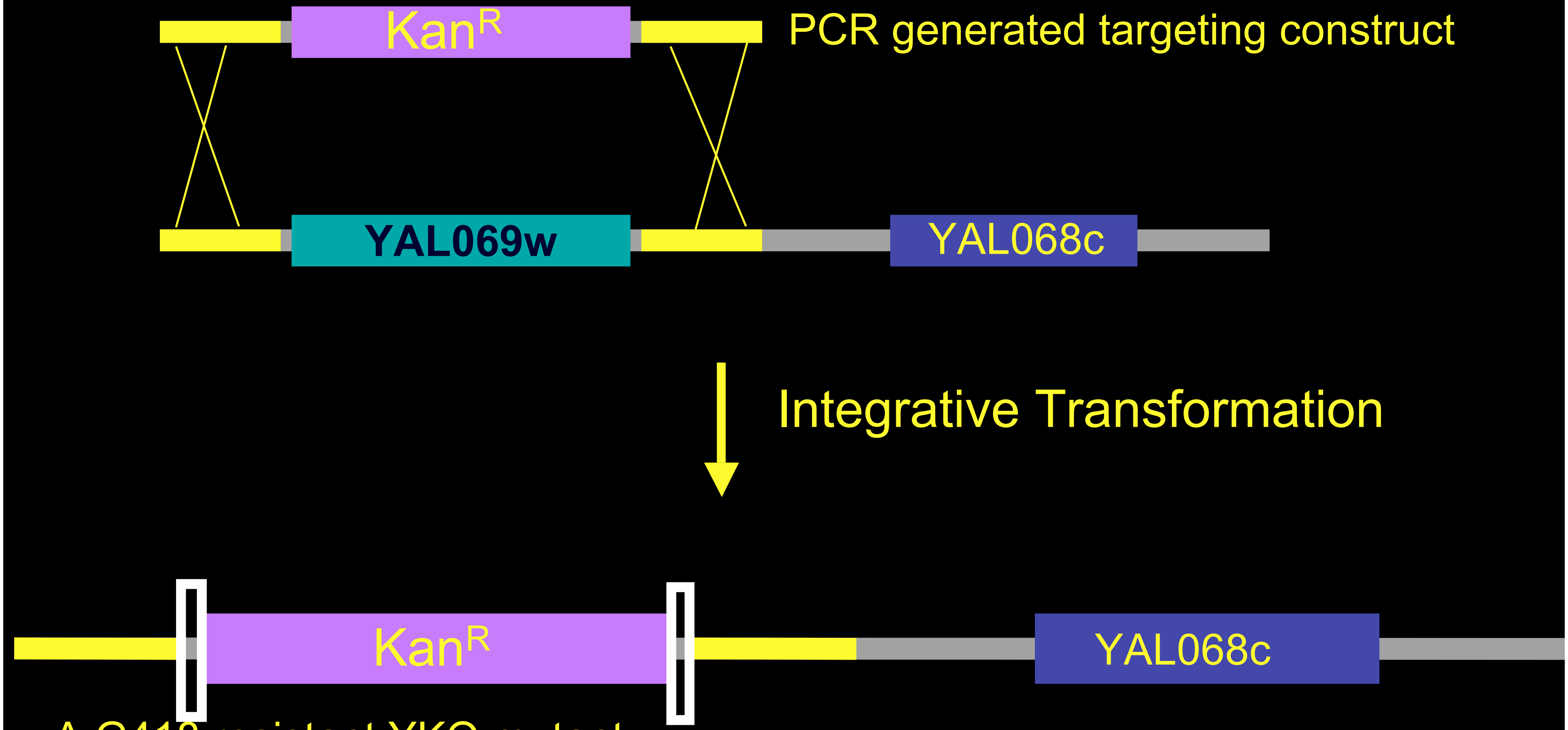


Combination with protein interaction map

Protein interaction map give “series circuits”
Genetic interaction map gives “parallel circuits”
Common data standards regarding gene/protein naming, etc. critical to network integration efforts



How it’s done: Yeast Knockouts (YKOs)

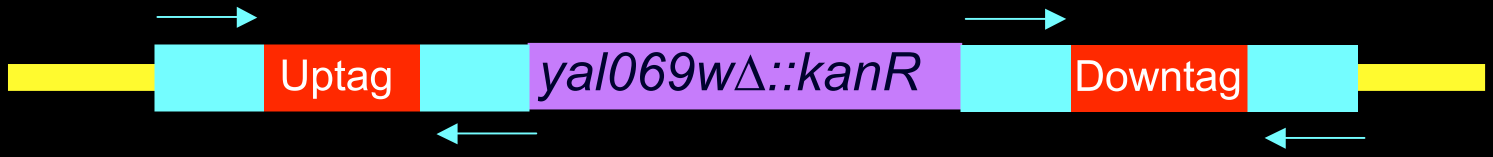


PCR generated targeting construct

Integrative Transformation

A G418 resistant YKO mutant
(*yal069w*Δ*::Kan^{R}*)

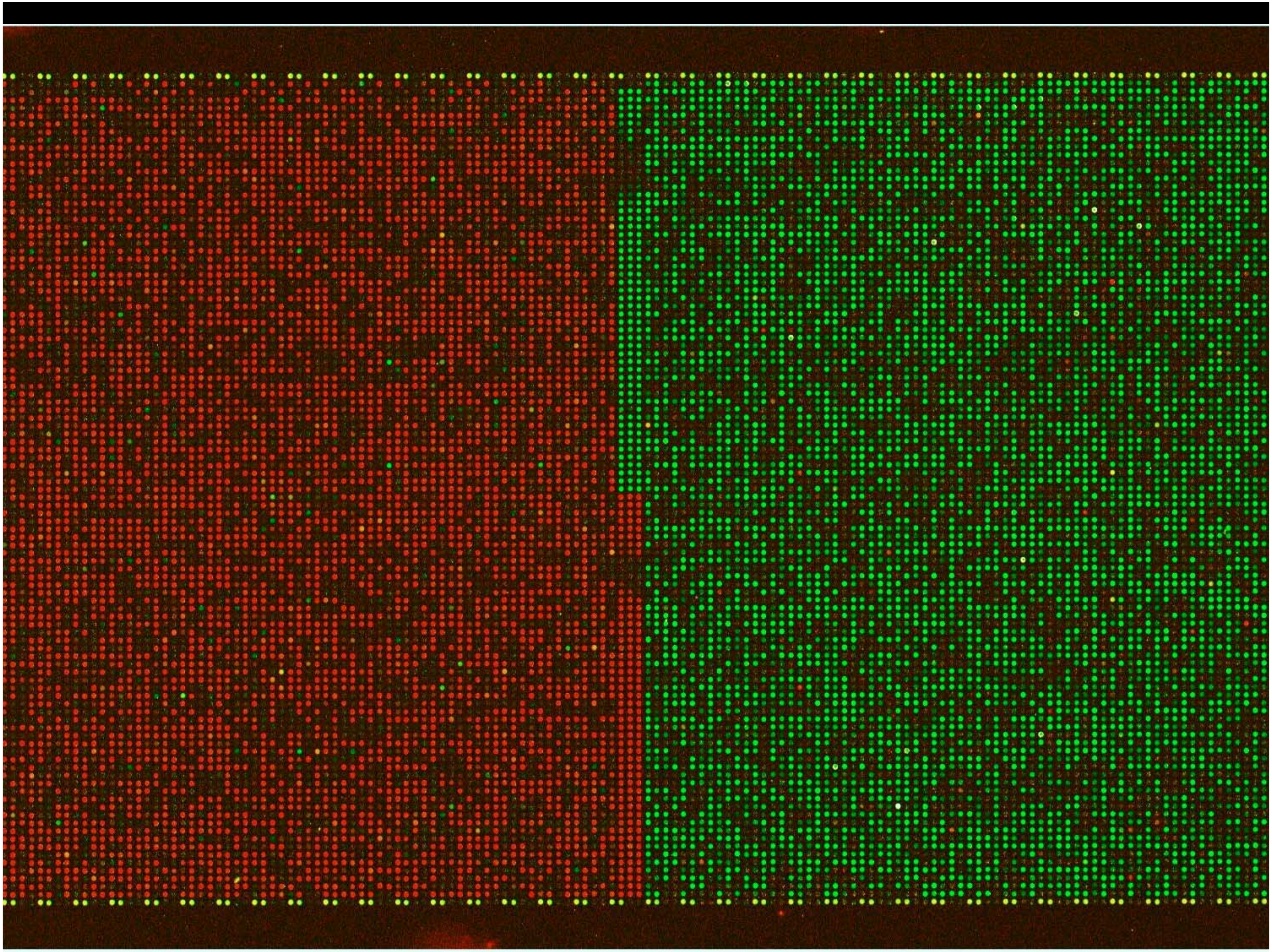
YKO features



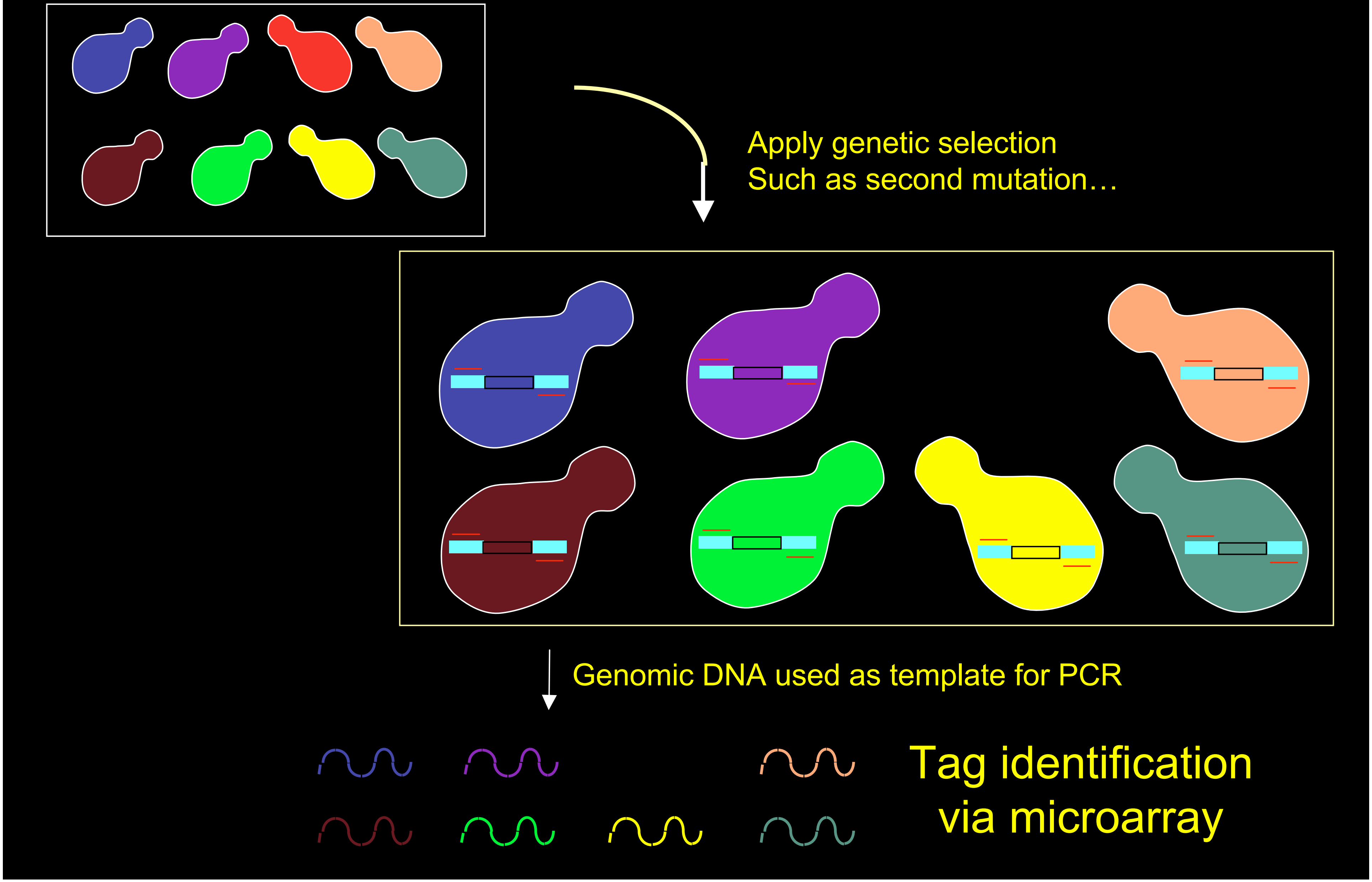
- Each YKO tagged with 2 unique sequences (20-mers) called UPTAGs and DOWNTAG

- UPTAGs and DOWNTAGs are flanked by universal priming sites

- 6200 yeast genes require 12400 *unique* TAG sequences that can also be put on microarrays (Shoemaker, Davis) and report on presence/absence of a given YKO in a complex mixture



TAGs allow parallel analysis of YKOs as a pool

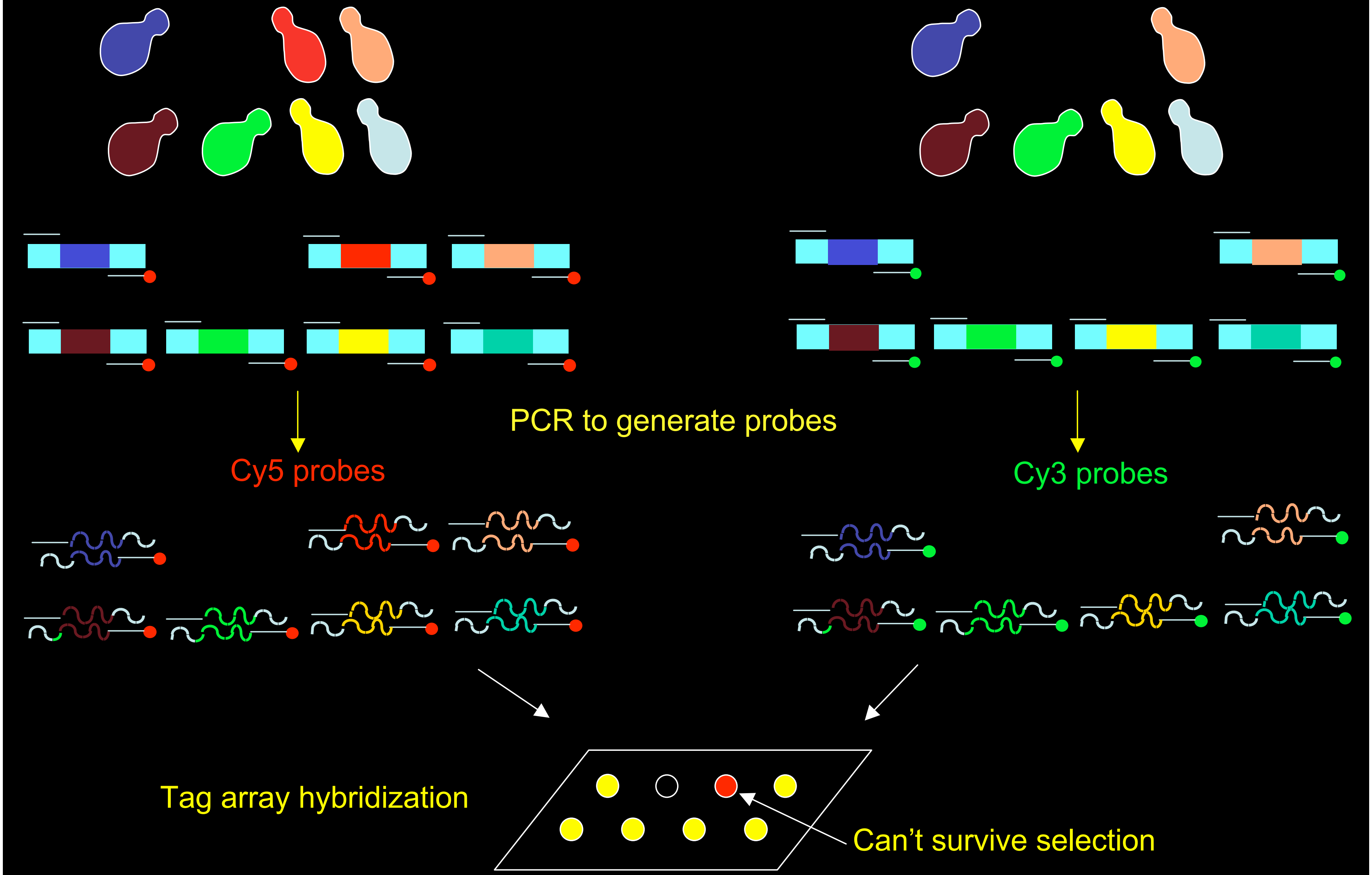


Apply genetic selection
Such as second mutation…

Genomic DNA used as template for PCR

Tag identification
via microarray

Parallel Analysis of YKO mutants continued…



PCR to generate probes

Cy5 probes

Cy3 probes

Tag array hybridization

Can’t survive selection

Keep in mind, from the microarrays we get a list of fluorescence intensities. At the end of the day we need to derive meaningful data in the form of

“It interacts”
Or
“It does not interact”

A number of problems relating to genetic properties of the strains used, and other methodological issues needed to be solved to make this method work well in practice. In particular, different mutants grow at different rates and have different transformation properties, leading to poor reproducibility.

Features of heterozygous diploid YKOs

- One wild-type copy and one deleted copy

- Problem: generally no detectable phenotype

- Best genetic quality compared to haploid and homozygous diploid YKOs; each YKO covered by wild-type copy- better data quality

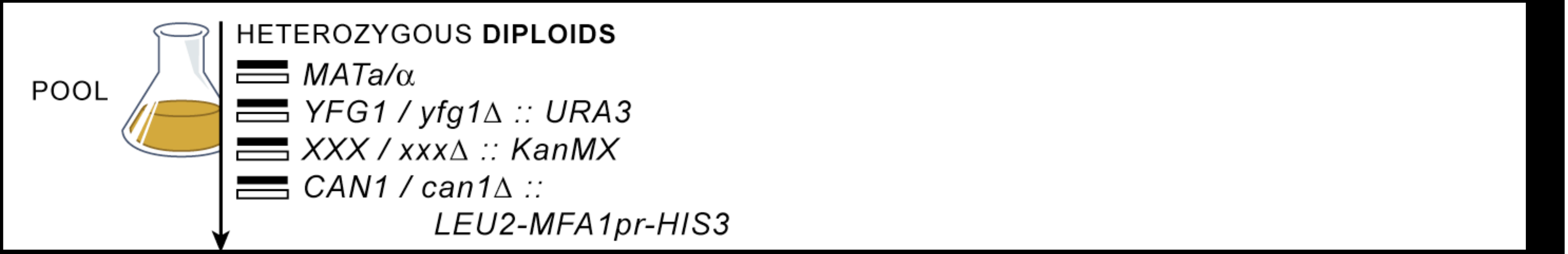
- YKOs behave ~uniformly; amenable to manipulation as pool - better data quality

- > Increased efficiency of integrative transformation (~10x) - better data quality

- Genetically manipulated pool stably maintained for later study (archive for validation/further studies)

- Use a “trick”: haploidization marker (Boone) uncovers mutation just prior to analysis

An experimental scheme for d-SLAM analysis



POOL

HETEROZYGOUS **DIPLOIDS**

MATa/α

YFG1 / yfg1Δ :: URA3

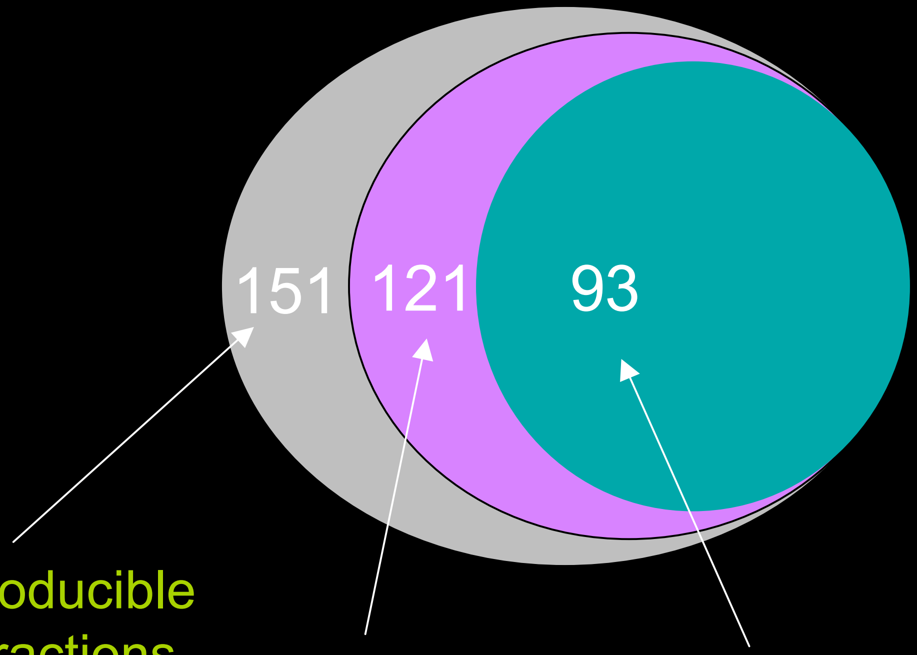
XXX / xxxΔ :: KanMX

CAN1 / can1Δ ::
LEU2-MFA1pr-HIS3

Validation

- Validations are done manually, one gene pair at a time, using a time-consuming random spore analysis
- This is our “gold standard” that allows false positives to be defined. By examining, for example, the top 100-200 hits, false positive rates can be determined
- False negatives are more problematic (carrying out a manual 1 X 6000 analysis is prohibitive) but rates can be estimated from known, previously determined interactor lists

Validating a d-SLAM screen: diploids are cleaner!



Reproducible Interactions By SLAM

Individually Confirmed as diploids

Confirmed by using the pre-existing haploid YKOs

- 121-93 =28 new interactions confirmable only in heterozygous diploid strains

Key question: do genes A and B interact or not?

For each gene pair analyzed, we would like to be able to make a simple statement (25,000,000 times)

It interacts (SL)
It interacts weakly (SF)
It does not interact
This cd be reduced to a binary outcome: interacts (SL/F), or not (a null value, for “no data” could also be provided)

However, we actually end up with a ranked list of interactors from the array experiments. Generally, there are no “clean breaks” in the ~continuous data allowing us to separate between the categories

This results in false positives and false negatives

Well-defined data quality metrics are critical

Such metrics must be systematic and rankable to be useful

Our goal (a work in progress) is to provide a statistical metric along the following lines:

Gene pair	Interact?	Confidence Score
A and B	yes	0.88

Factors that could influence the Confidence Score and how…

Factor	Effect on CS
Known to interact by indiv biological test	CS=1.0
High C/E ratio	increase
Query finds target AND Target finds query	increase
UPTAG and DNTAG agree	increase
UPTAG or DNTAG probe “noisy”	decrease
Slow growing query gene mutant	decrease
Slow growing target mutant	decrease

Algorithms for determining predictive confidence scores can and will be tested/trained on validated samples empirically

Best datasets have well-defined validation controls

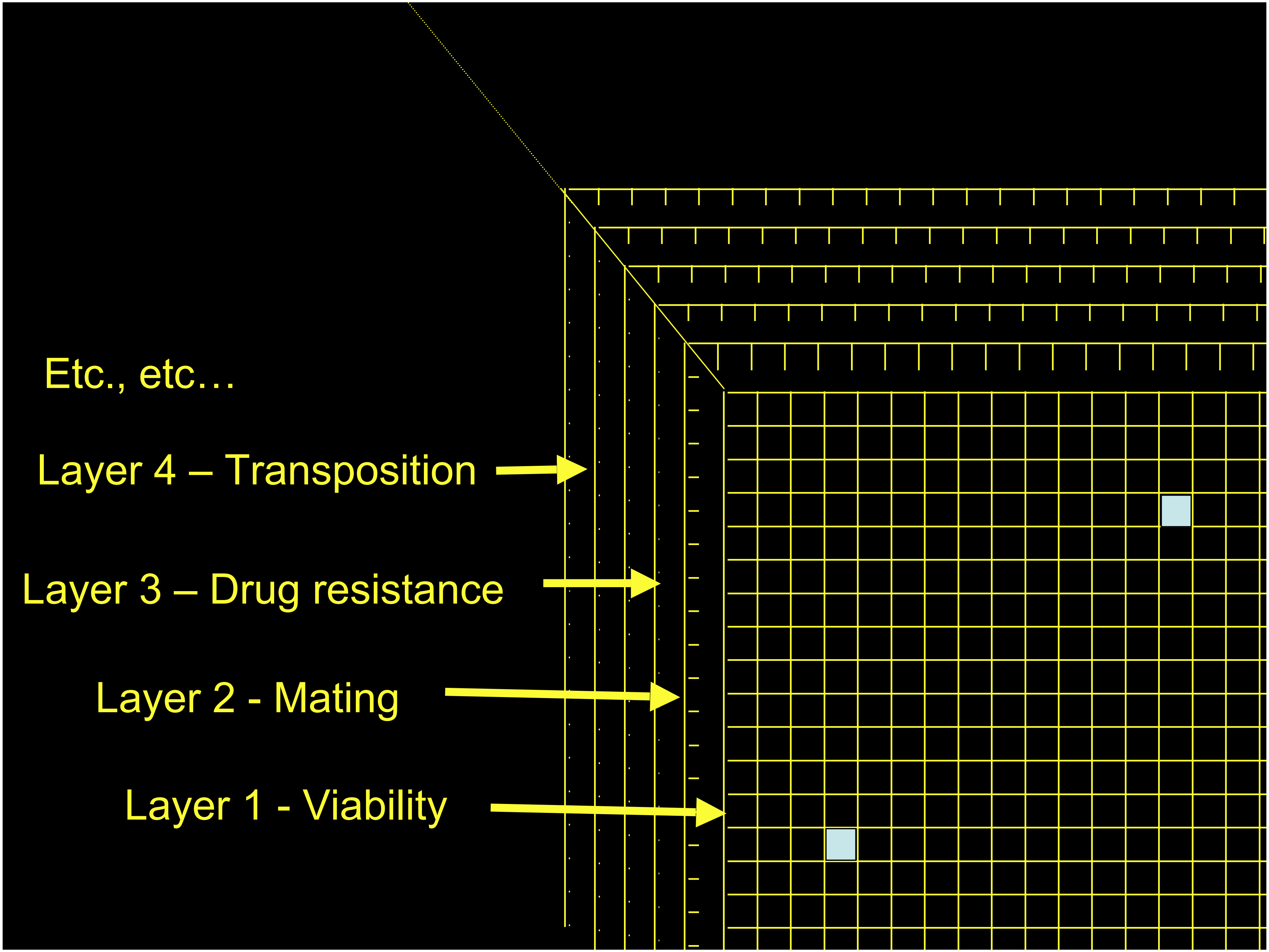
- Populations of mutants made as artificial mixtures
- Well characterized query gene experiment done in triplicate and compared to biologically validated final list

- Data should be provided at different levels of abstraction for different levels of use
- High level --- Simple binary interaction partner lists, with confidence scores attached
- Medium level -- All factors influencing confidence score calculation tabulated
- Raw data level -- Raw array images or .GPR files

Genome SLAM will generate an interaction map

However, this 5000 X 5000 matrix will be only the first layer of a growing, 3D interaction database

	Yfg1	Yfg2	Yfg3	Yfg4	Yfg5	Yfg6	Yfg7	Yfg8	Yfg9	Yfg10	Yfg11	Yfg12	Yfg13	Yfg14	Yfg15	Yfg16	Yfg17	Yfg18	Yfg19	Yfg20	Yfg21
Yfg1																					
Yfg2																			■		
Yfg3																					
Yfg4														■							
Yfg5																					
Yfg6																					
Yfg7																					
Yfg8																					
Yfg9																					
Yfg10																					
Yfg11																					
Yfg12																					
Yfg13																					
Yfg14				■																	
Yfg15																					
Yfg16																					
Yfg17																					
Yfg18																					
		■																			



Etc., etc…

Layer 4 – Transposition

Layer 3 – Drug resistance

Layer 2 - Mating

Layer 1 - Viability

The ultimate protein-related property in need of a standard(s): Mutant phenotype

- Phenotypes are incredibly diverse
- The existence of large systematic mutation collections like the YKO collection means large systematic datasets on phenotype are being captured by many researchers
- These datasets provide incredibly rich source of information on protein function and thus are a treasurehouse of knowledge about proteins
- These datasets are recorded unsystematically and not centrally databased
- No standard data types currently exist for phenotypes except for some very simple ones, like viable/inviable
- We aim to build a “Phenotypes Database” or PhD precisely to warehouse this type of information, in conjunction with SGD

“Genestrings” -- the concept behind PhD

- Genetic screens yield information about *sets* of genes
  - –Survival, growth rates, cell morphology
- Data for each gene are reduced to a single character
  - –‘0’ or ‘1’ for boolean values, ‘0’ to ‘9’ for log P-values
- Sets of characters are represented as strings of 6000 chars
  - –character 1 = SGD S000000001 = YAL001C
- Why strings?
  - –Strings are highly compressible and portable

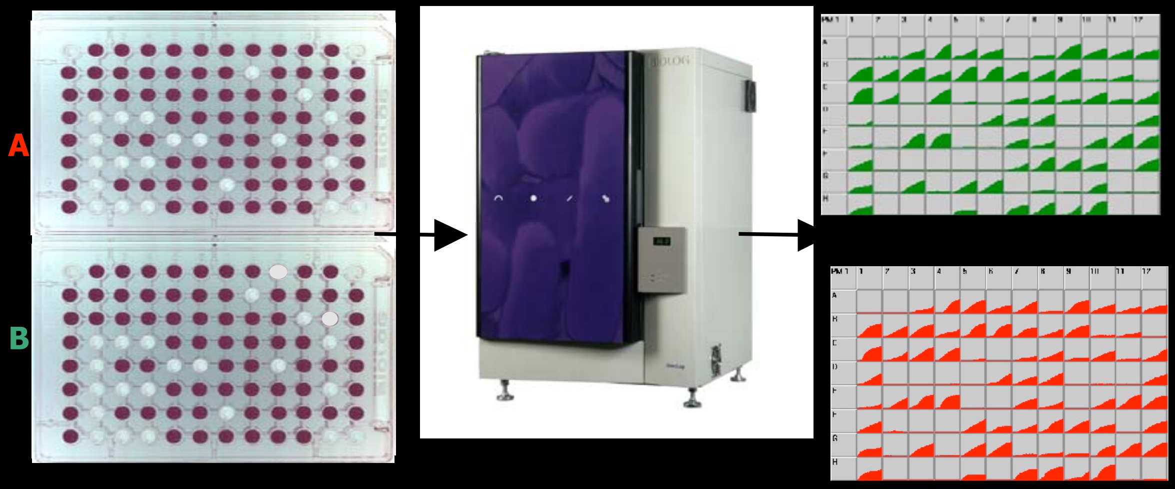
Daniel Yuan, Ph.D.

Advantages of Genestrings

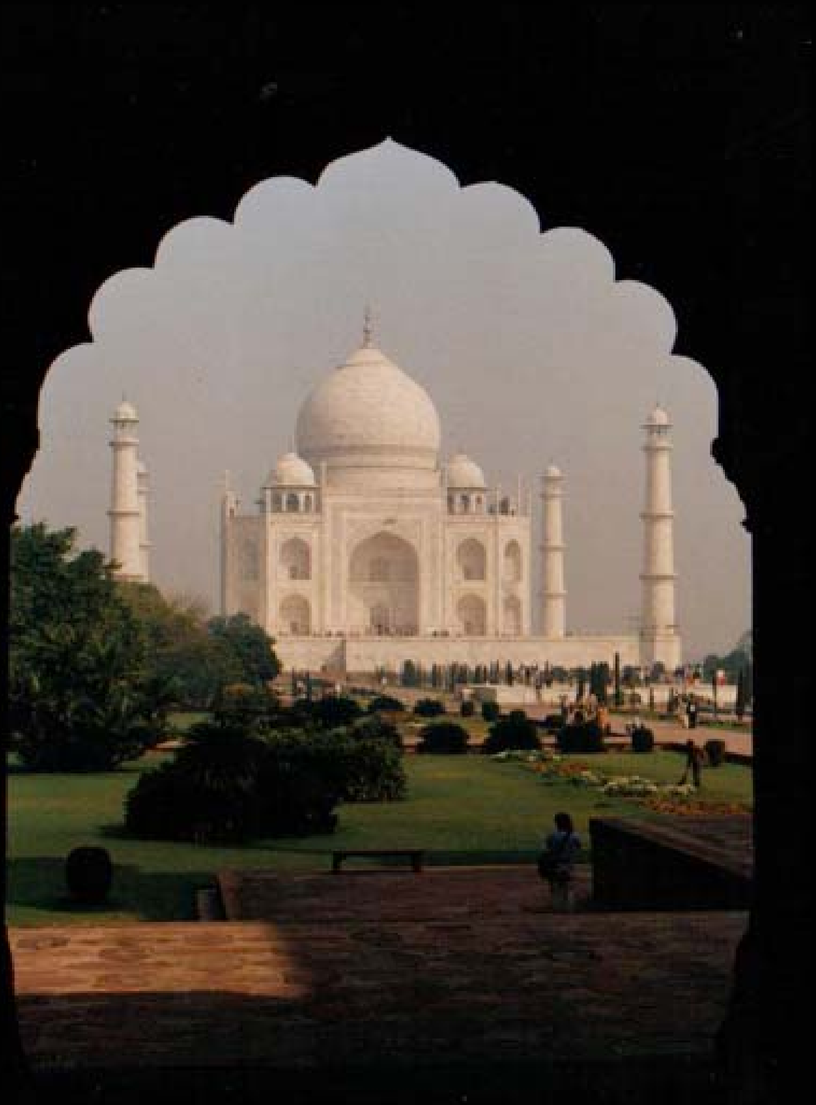
- Currently, lists of genes:
  - are constructed ad hoc (gene names, tabular formats)
  - are cumbersome to compare (order not defined)
- By contrast, genestrings:
  - have unambiguous structure
  - are computer-ready
  - include semiquantitative and categorical data
  - represent data for 6000+ genes in a few lines of text

High throughput phenotyping - easier standardization

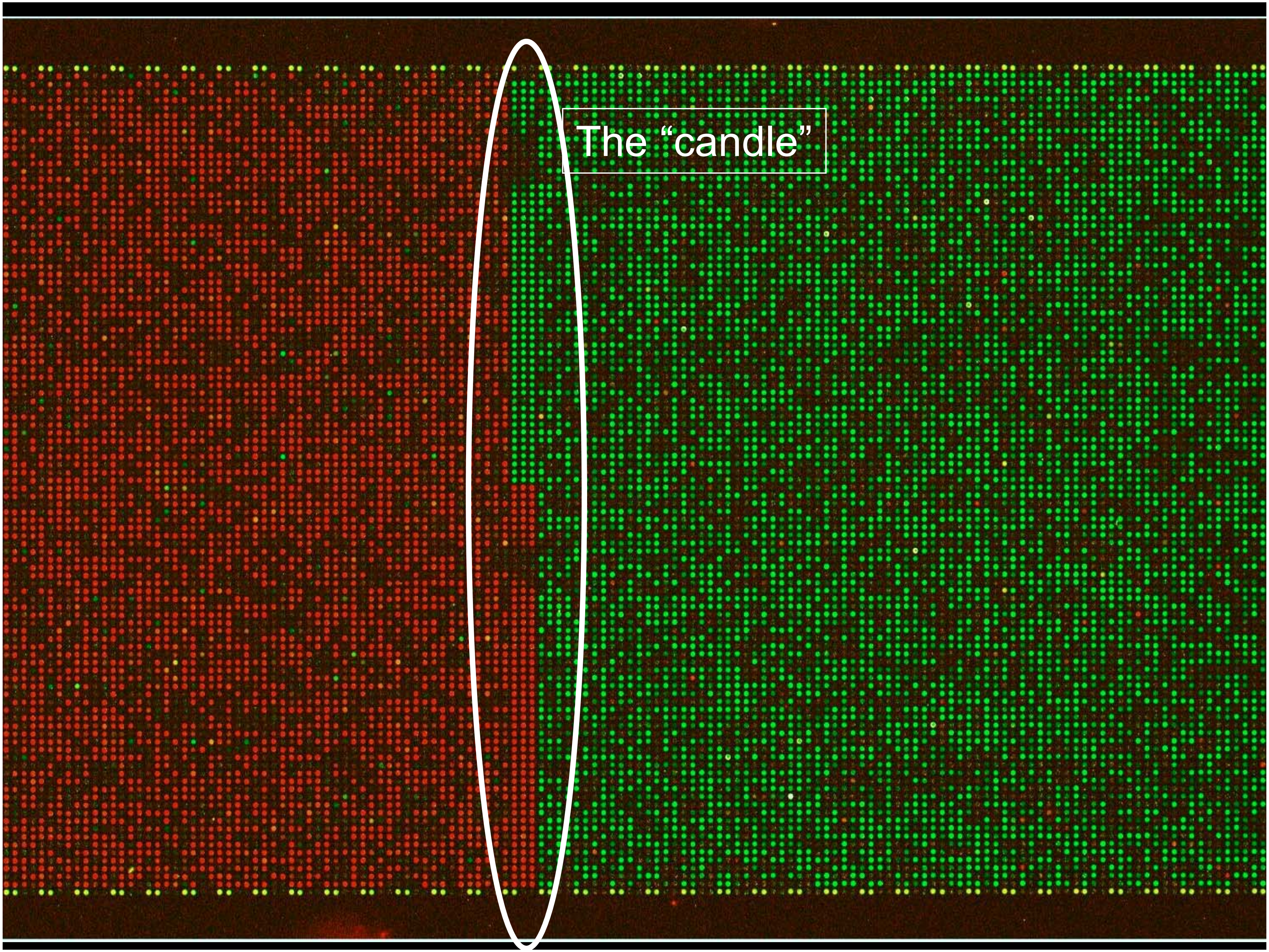
We are working with Biolog, a company with a platform for systematically collecting 2000 phenotypes at a time



This is one solution to the phenotype standard problem but it only addresses one class of phenotypes- growth - and is too expensive a technology for an average lab



After Jan 4, 2005…



The “candle”