

Combinatorial Pooling Enables Selective Sequencing of the Barley Gene Space

Presented by

Stefano Lonardi

Computer Science and Engineering
University of California Riverside

&

Timothy Close

Botany and Plant Science
University of California Riverside



Hosted by
Shawn Yarnes
Plant Breeding and Genomics

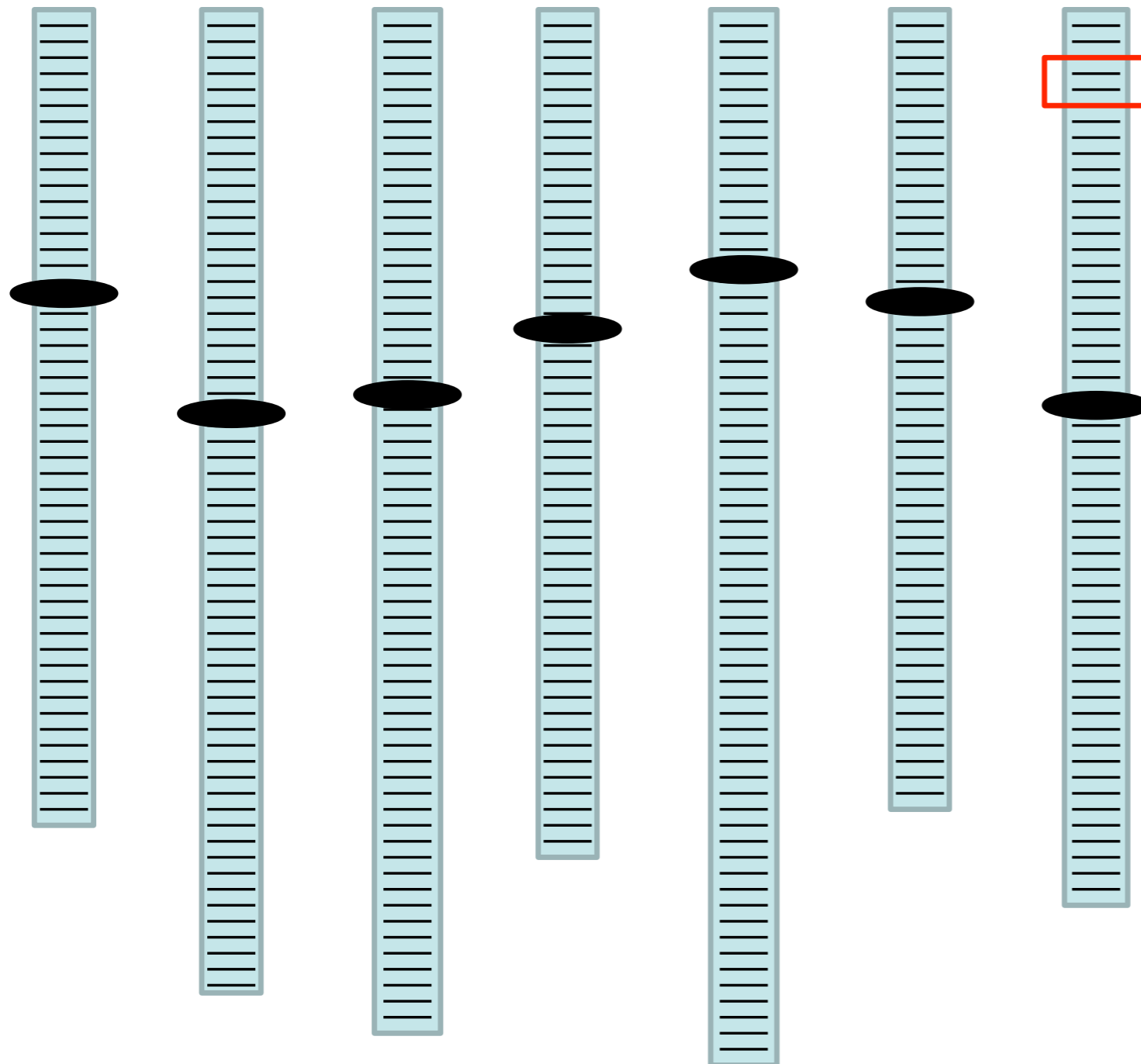


Barley genome (*H. vulgare*)

- Diploid
- Seven chromosomes
- Size is ≈ 5.3 Gb
 - $\approx 36\times$ the size of *Arabidopsis*
 - $\approx 12\times$ the size of rice
 - $\approx 9\times$ the size of cowpea
- Highly repetitive (>90%)
- Genome too repetitive for complete WGS from short reads

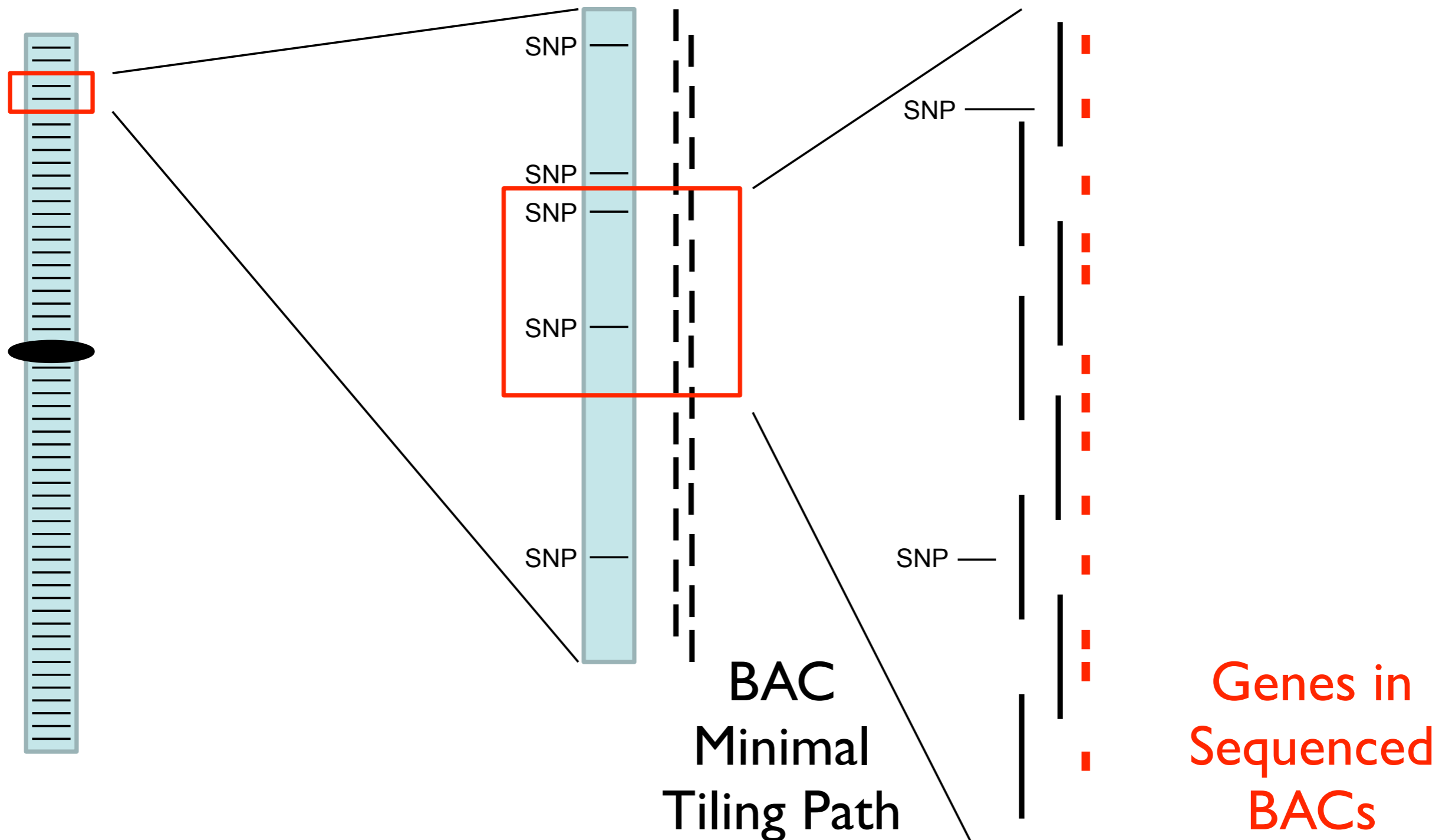


Location of a Trait on a Genetic Map



Trait position:
“What candidate
genes are in this
region?”

Location of a Trait on a Genetic Map



Barley genome (*H. vulgare*)

- BAC (Bacterial Artificial Chromosome) a 100-150kb fragment of the target genome propagated in *E.coli*
- Genes are not distributed evenly along the genome: they are clustered in gene-rich regions, thus a BAC carrying one gene is likely to carry several genes
- Strategy (*selective sequencing*)
 - Identify gene-enriched BACs
 - Build an overlap map (*physical*) for these BACs
 - Sequence a minimally redundant subset (*minimal tiling path; MTP*)

BAC-by-BAC vs. WGS

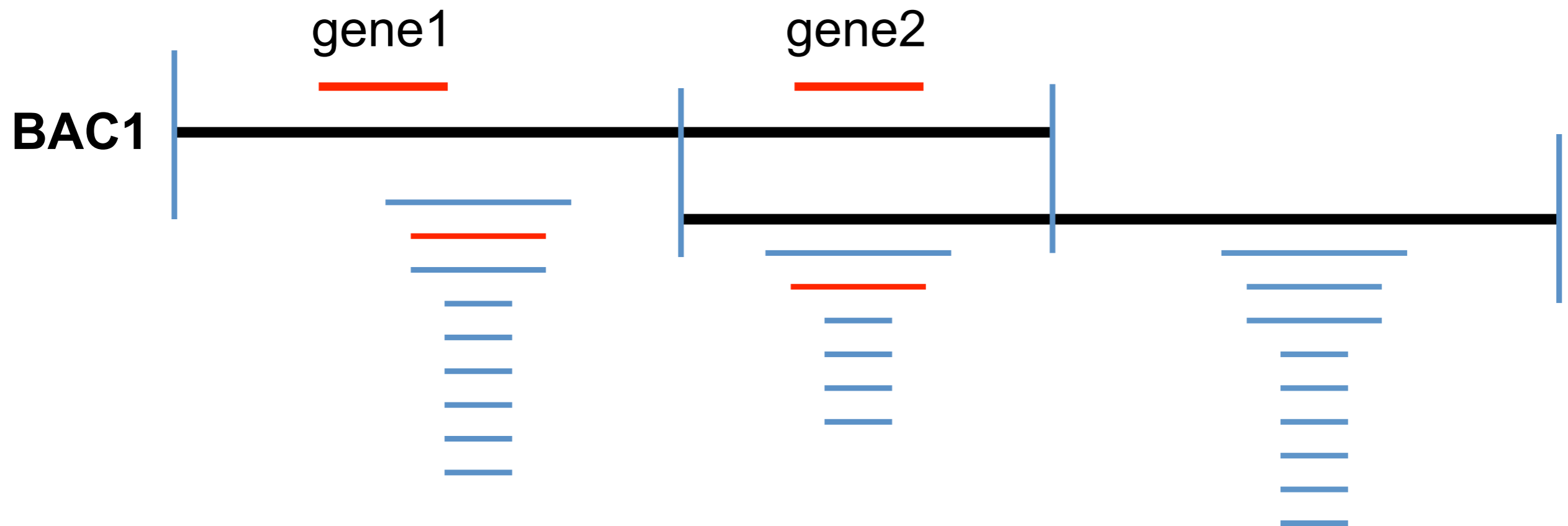
- Pros

- Can be selective (*i.e.*, gene enrichment)
- Work can be distributed across several labs
- Assembly can be carried out BAC-by-BAC (helps dealing with high repeat content)

- Cons

- Need BAC library & overlap map (*physical map*)
- *E. coli* contamination in BAC DNA
- Need to handle large number of individual samples

Outcome is sets of unordered sequences
allocated to bins defined by MTP BAC ends



Barley BAC physical map

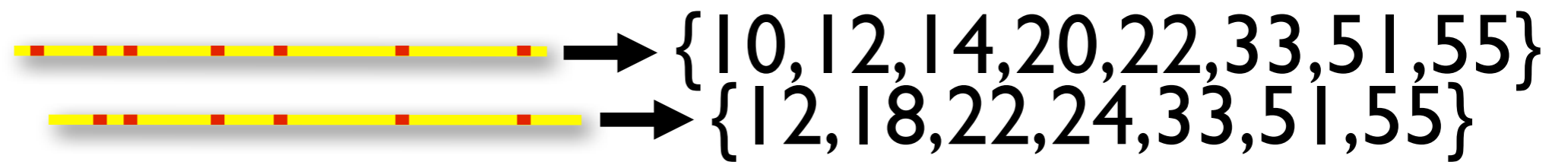
- Started from 6.64x genome equivalent BAC library for Morex barley (313,344 BACs) [Yu *et al.*, 2000]
- Selected 83,831 gene-positive BACs, then fingerprinted using HICF (five restriction enzymes)

DNA Fingerprinting



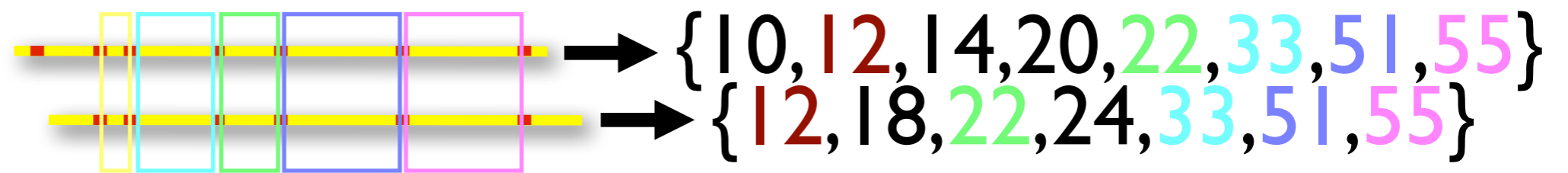
- BACs are “digested” with restriction enzymes that cut DNA at specific sites

DNA Fingerprinting



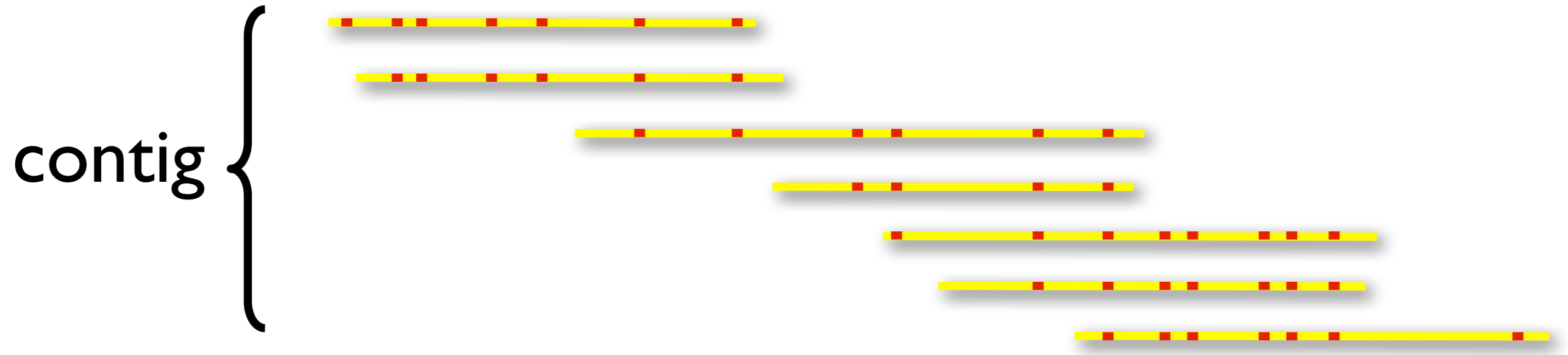
- The length of the fragments obtained after digestion are measured

DNA Fingerprinting



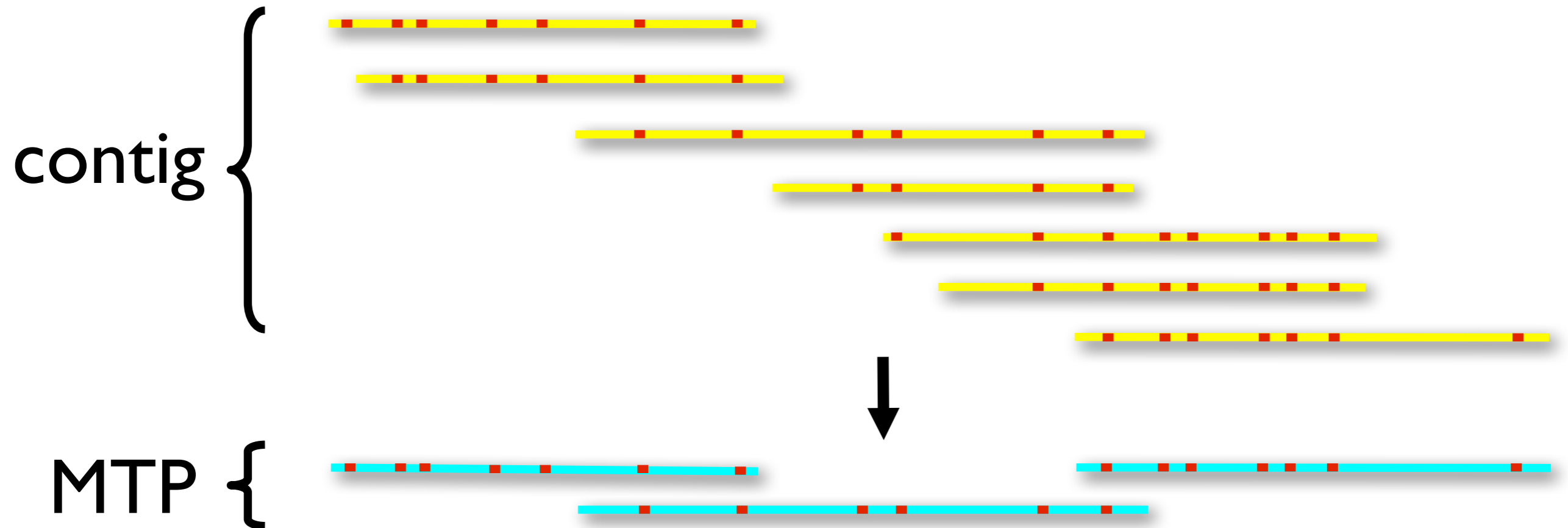
- Two BACs are declared overlapping if they share a large number of common “lengths”

DNA Fingerprinting



- A set of overlapping BACs is a *contig*

Minimum Tiling Path (MTP)



- 15,720 BACs were identified as *minimal tiling path* (MTP) clones, for a total of ~1,700 Mb [Bozdag et al., *Proc. WABI* 2008]

Next-Generation Sequencing

- NGS instruments have a fixed number of ‘lanes’ for DNA samples (e.g., Illumina has 8)
- Allocating one BAC to each individual lane would be expensive and wasteful
- Need to “multiplex” many BACs on the same lane, but DNA barcoding does not scale readily to hundreds or thousands of samples

Combinatorial Pooling

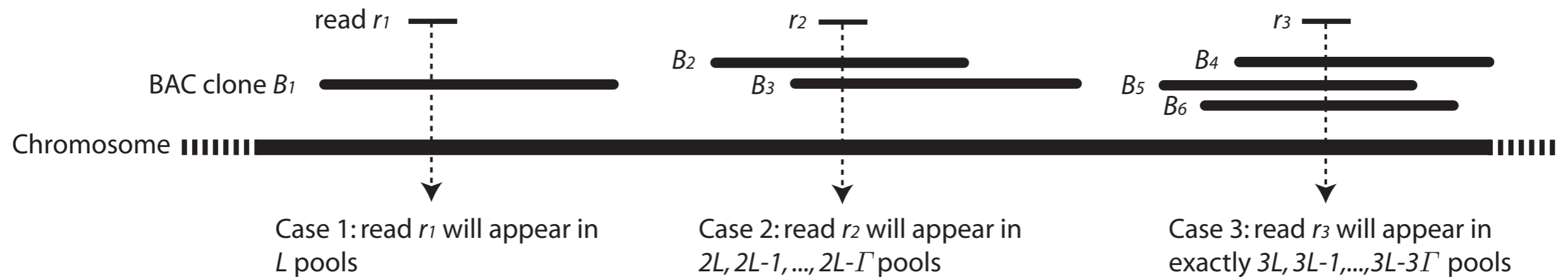
- *Idea*: Replicate each BAC in a set of pools according to a *combinatorial pooling* scheme so that the identity of a BAC is encoded in the pattern of pools (*signature*) where it is contained

[by transitivity, corresponding sequence reads will exhibit the same pool pattern]

Combinatorial Pooling

- A *shifted transversal* design is defined by (P, L, T, d) such that P is a prime, $P^{T+1} \geq N$ and $\text{floor}[(L-1)/T] \geq d$ [Thierry-Mieg, *BMC Bioinfo* 2006]
- Properties
 - Number of pools is PL
 - *Decodability* is d
 - A BAC is replicated in L pools
 - Each pool contains P^T BACs
 - Two BACs can share at most T pools

Need a 3-decodable design



Set $L=7, \Gamma=2 \Rightarrow$ 3-decodable

Several 3-decodable 7-layer designs

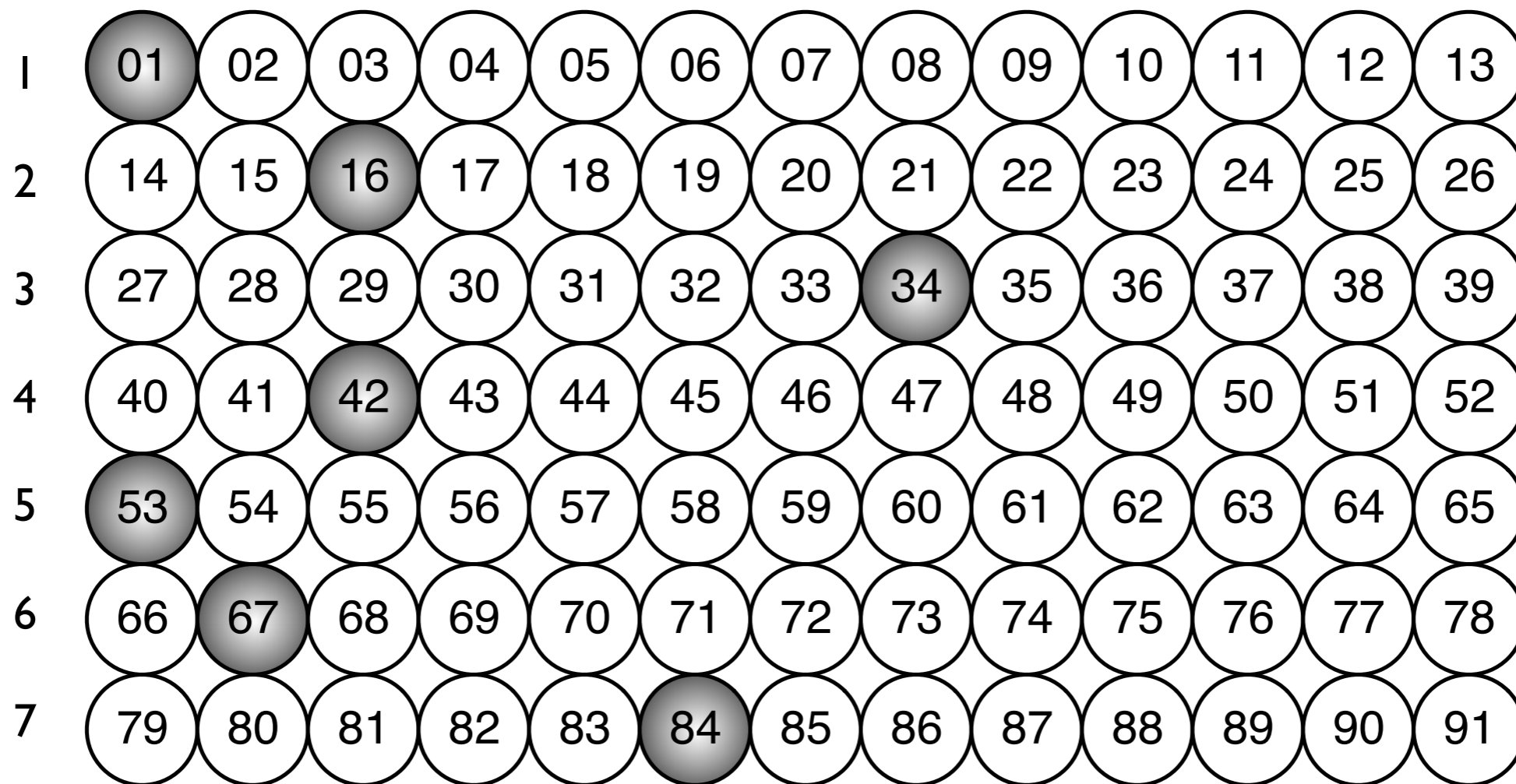
P	BACs/pool (P^2)	Total BACs (P^3)	Total pools ($7 \times P$)	<u>Total BACs</u> Total pools
7	49	343	49	7.0
11	121	1,331	77	17.3
13	169	2,197	91	24.1
17	289	4,913	119	41.3
19	361	6,859	133	51.6
23	529	12,167	161	75.6
29	841	24,389	196	124.4

Pooling design and sequencing

- We divided the 15,720 barley MTP BACs in
 - seven sets (Hv3-Hv9) of 2,197 BACs pooled according to the ST design ($P=13, L=7, T=2, d=3$)
 - one set (Hv10) of 1,331 BACs pooled according to the ST design ($P=11, L=7, T=2, d=3$)
- Each set of 91 pools run on one Illumina flowcell: each of the seven available lanes was assigned 13/16/20 pools multiplexed via DNA-barcoding (*via* custom adapters)

Encoding BAC signatures

Layer



BAC signature

{01, 16, 34, 42, 53, 67, 84}

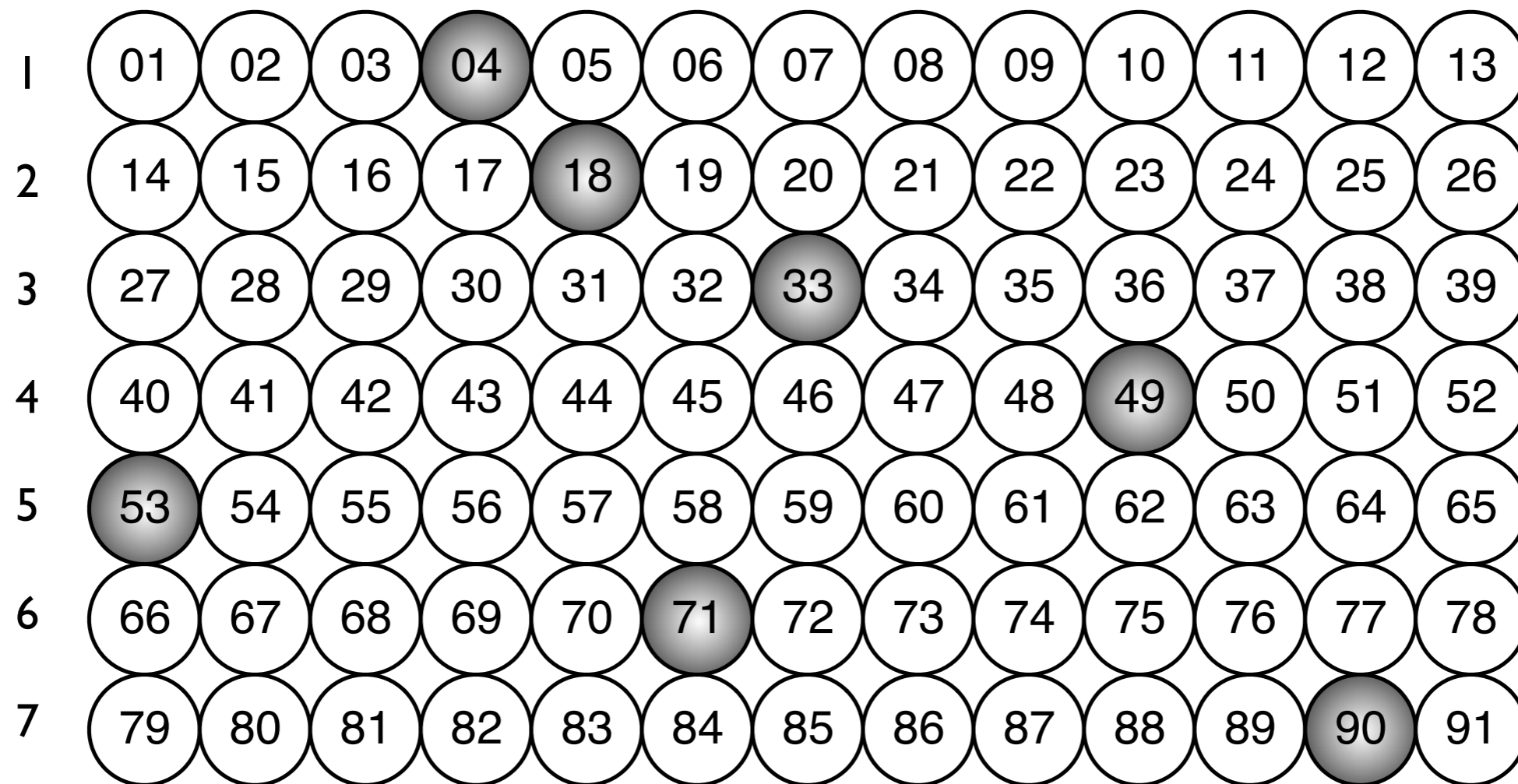
BAC

#0001

- 2197 BACs
- 91 pools: 7 layers, 13 pools per layer
- 169 BACs per pool
- Each BAC in 7 pools, one per layer

Encoding BAC signatures

Layer



BAC signature

{04, 18, 33, 49, 53, 71, 90}

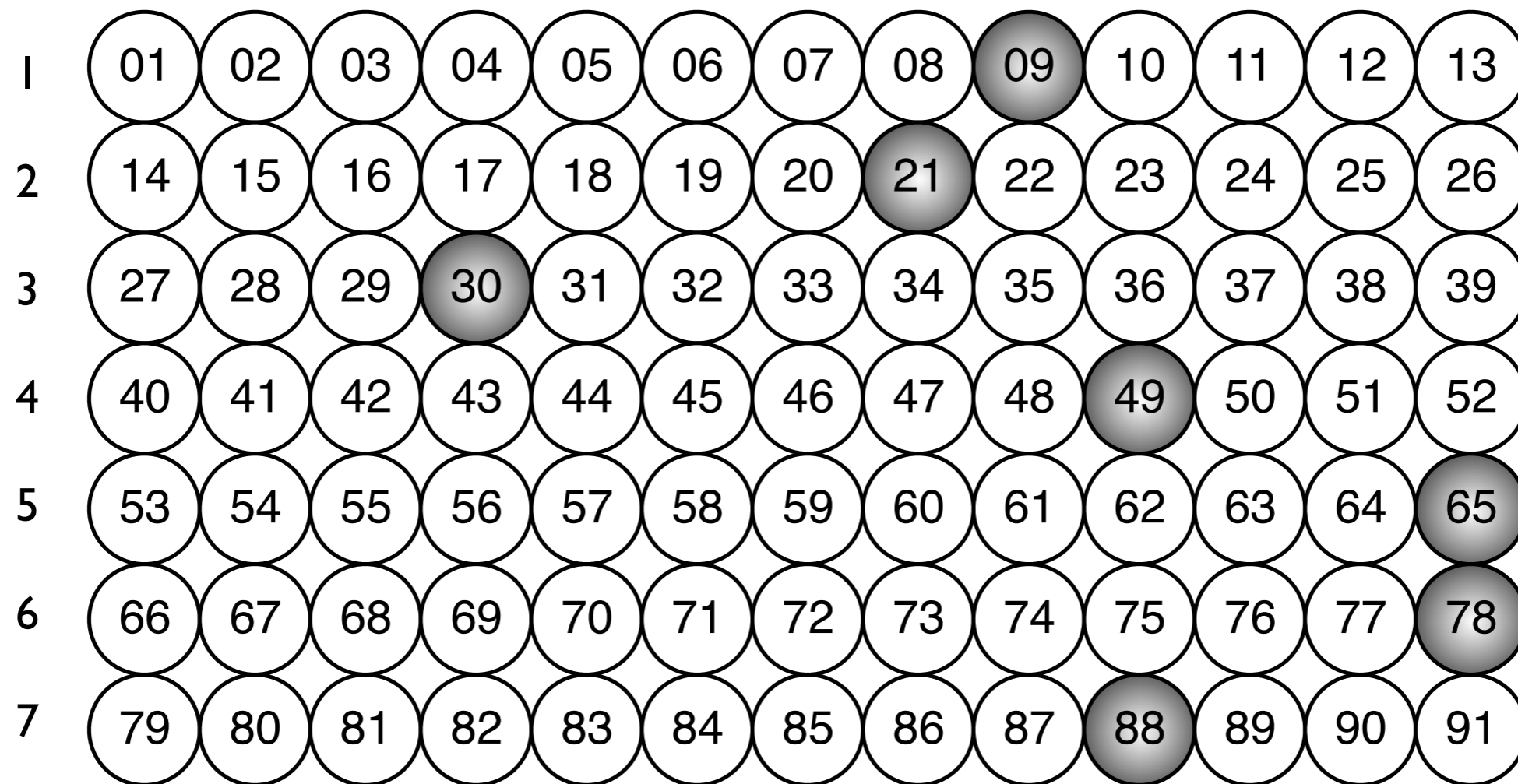
BAC

#0002

- 2197 BACs
- 91 pools: 7 layers, 13 pools per layer
- 169 BACs per pool
- Each BAC in 7 pools, one per layer

Encoding BAC signatures

Layer



BAC signature

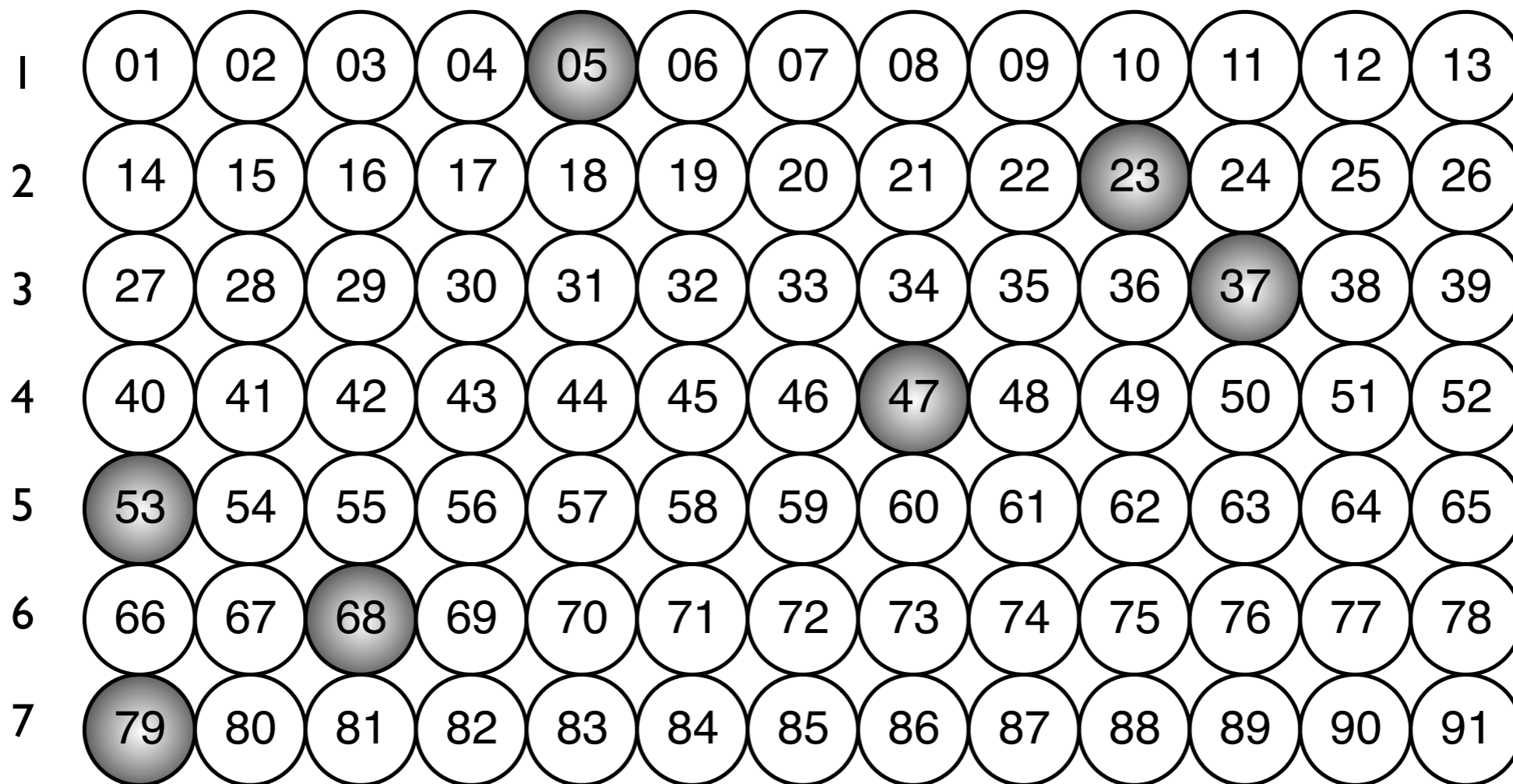
{09, 21, 30, 49, 65, 78, 88}

BAC

#0003

- 2197 BACs
- 91 pools: 7 layers, 13 pools per layer
- 169 BACs per pool
- Each BAC in 7 pools, one per layer

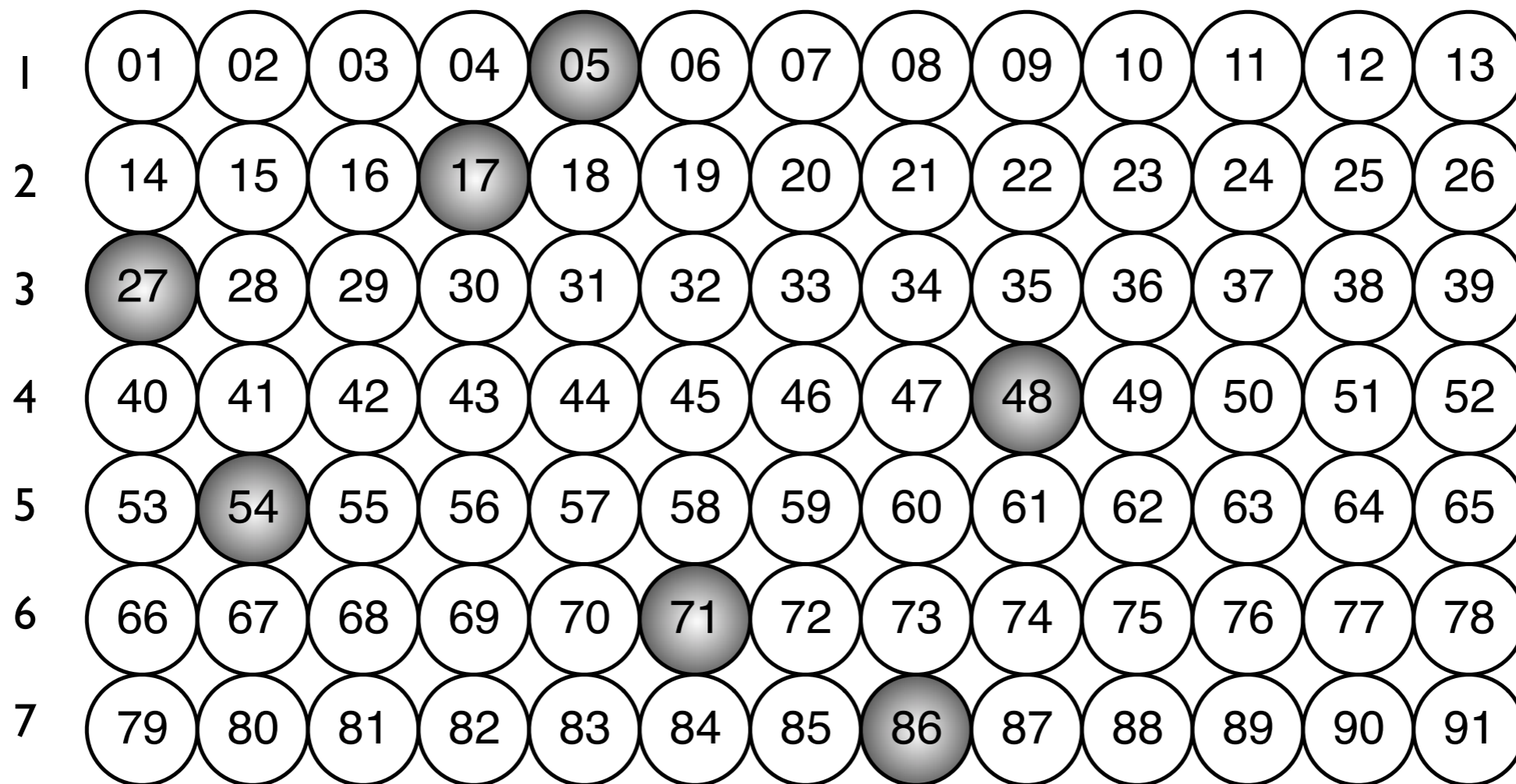
Encoding BAC signatures



... and so on for all 2,197 BACs ...

Decoding read signatures

Layer

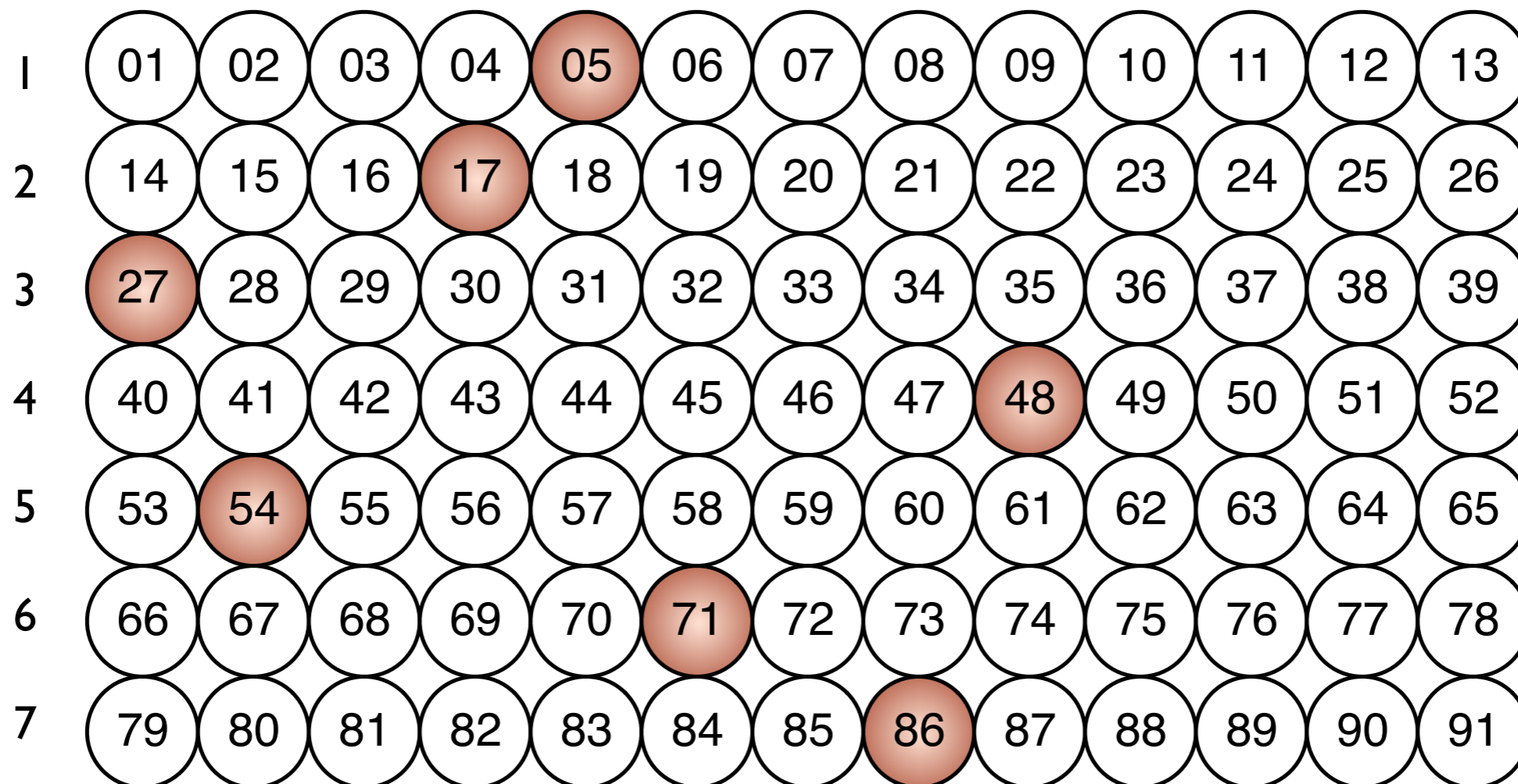


Read signature

{05, 17, 27, 48, 54, 71, 86}

Decoding read signatures

Layer



BAC signature

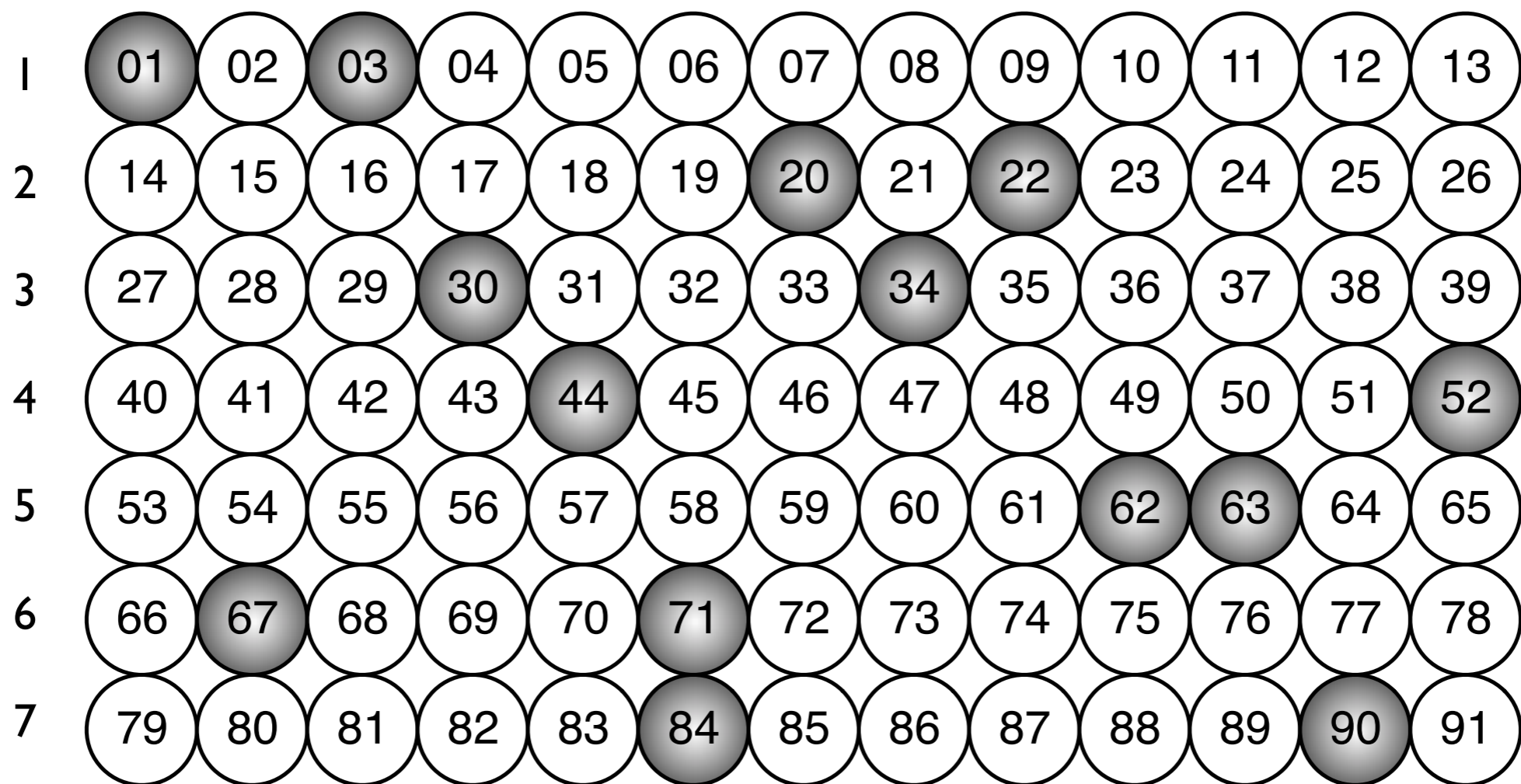
{05, 17, 27, 48, 54, 71, 86}

BAC

#0006

Decoding read signatures

Layer

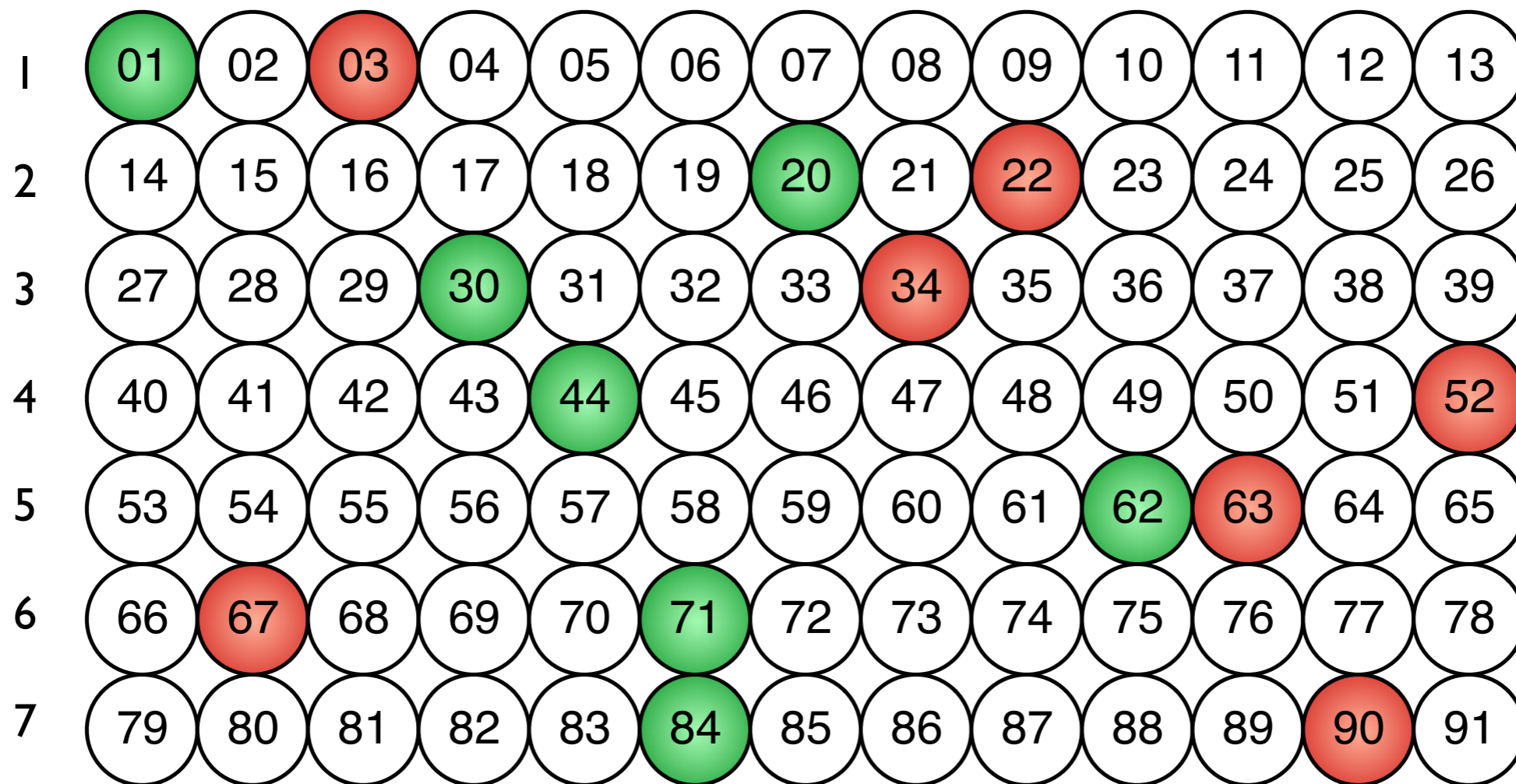


Read signature

{01, 03, 20, 22, 30, 34, 44, 52, 62, 63, 67, 71, 84, 90}

Decoding read signatures

Layer



BAC signatures

{03, 22, 34, 52, 63, 67, 90}

{01, 20, 30, 44, 62, 71, 84}

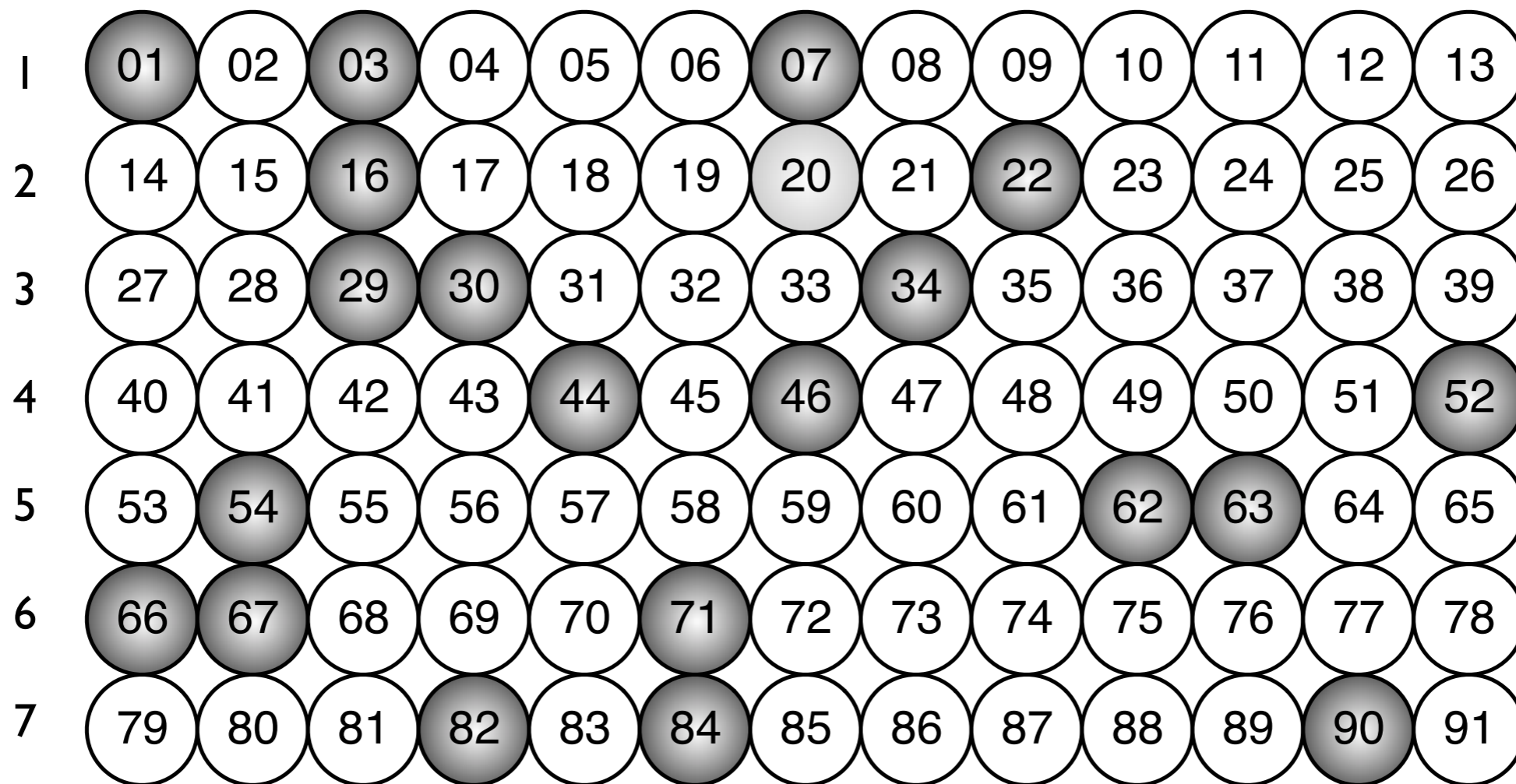
BAC

#0296

#1179

Decoding read signatures

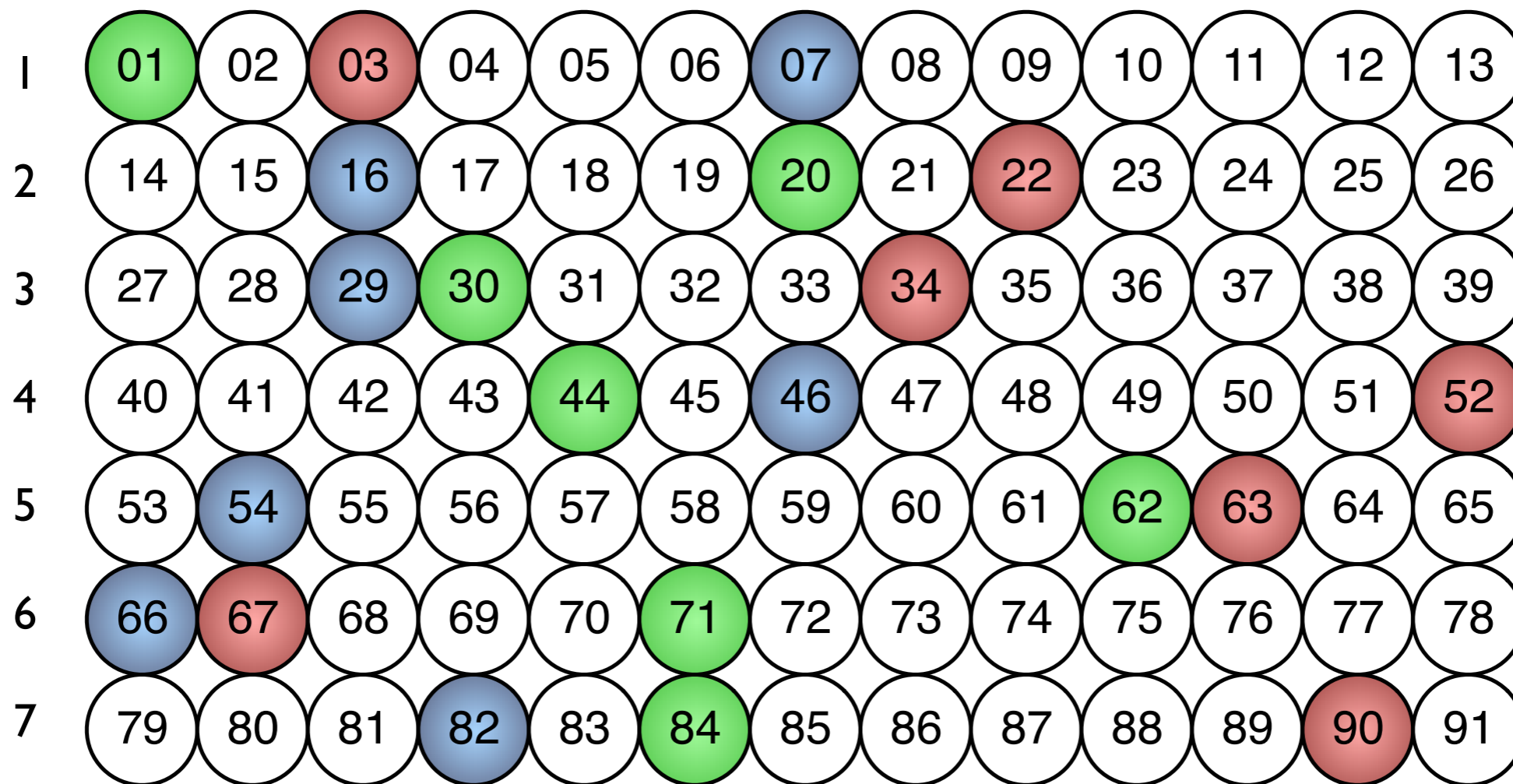
Layer



Read signature {01, 03, 07, 16, 20, 22, 29, 30,
34, 44, 46, 52, 54, 62, 63, 66, 67, 71, 82, 84, 90}

Decoding read signatures

Layer



{03, 22, 34, 52, 63, 67, 90}

#0296

{01, 20, 30, 44, 62, 71, 84}

#1179

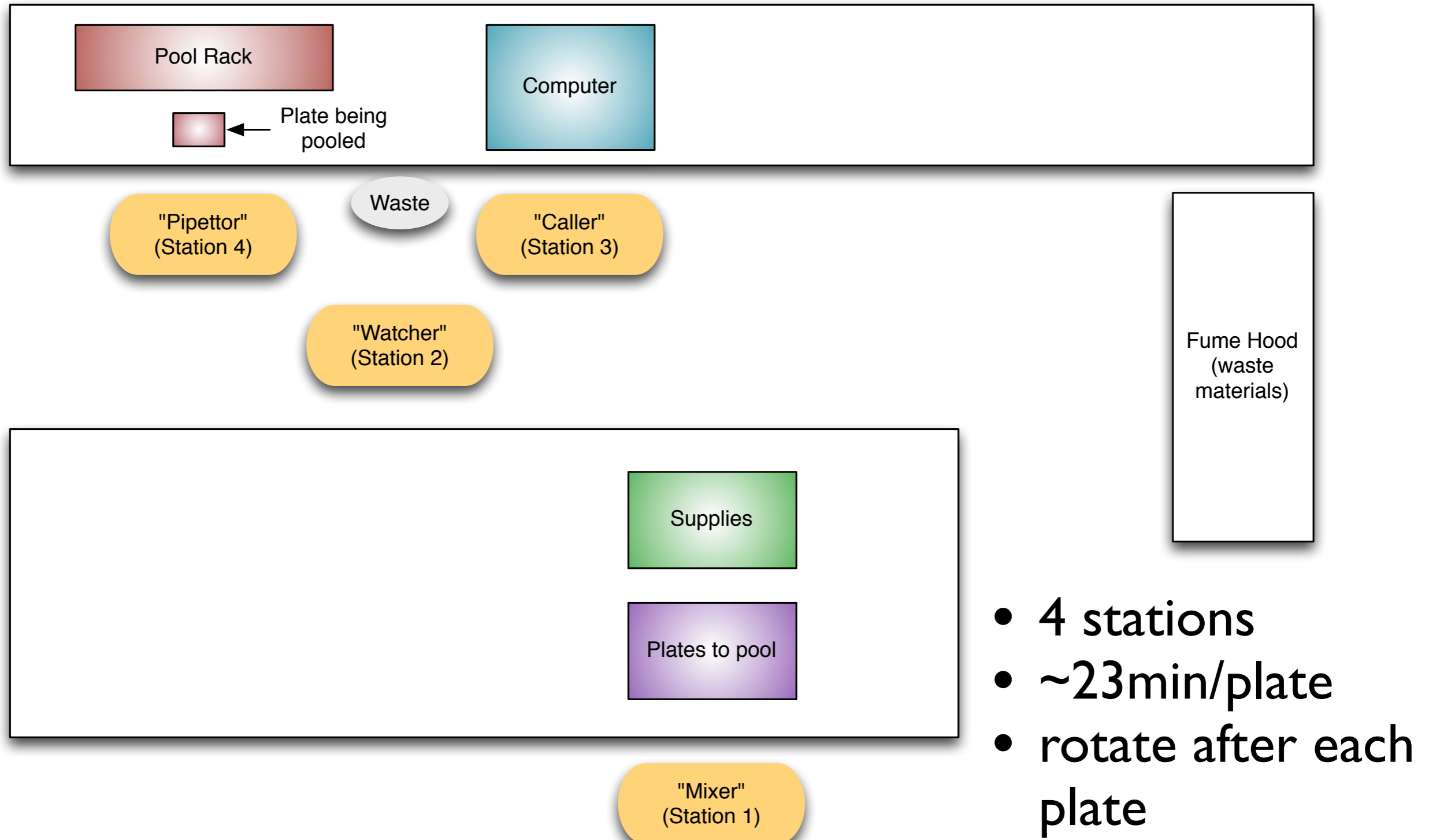
{07, 16, 29, 46, 54, 66, 82}

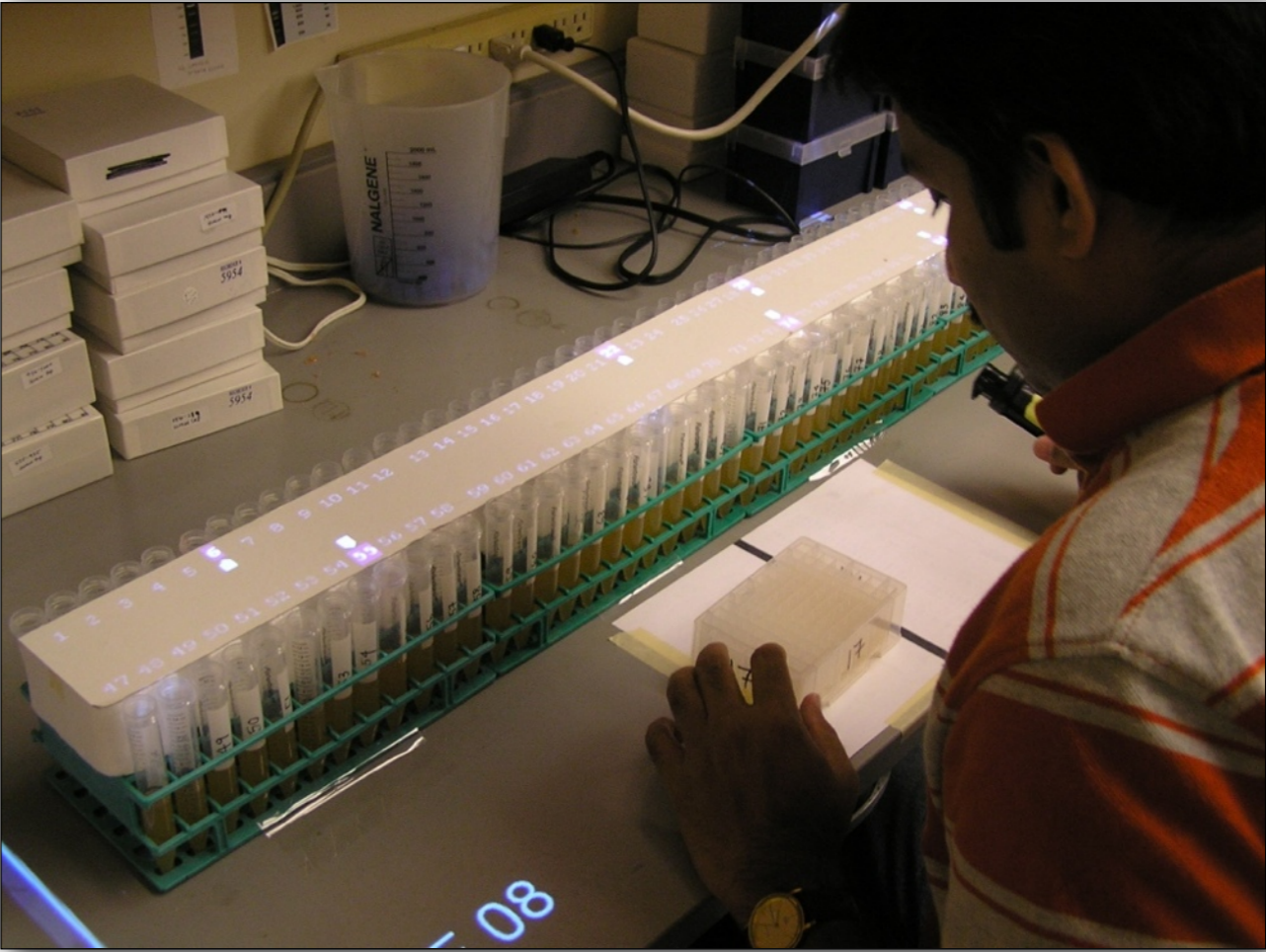
#1861

Decoding/deconvolution problem

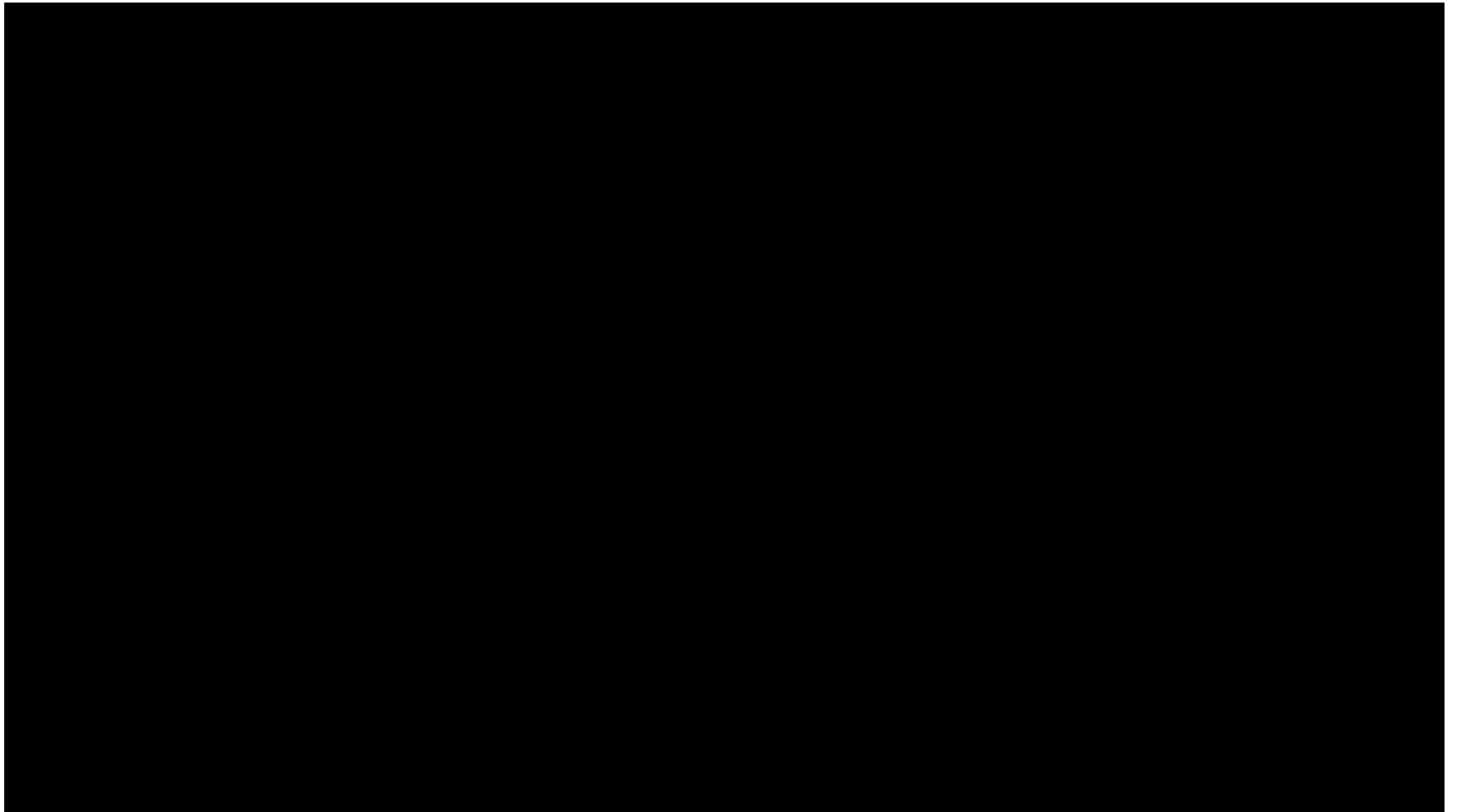
- *Input:* given a set of 91 pools of reads, and the signatures of 2,197 BACs
- *Output:* an assignment of each read to 1, 2 or 3 BACs
- *Challenge:* number of input reads is in the hundreds of millions; need an accurate time- and memory-efficient method
- *Software tool:* HashFilter
(<http://www.cs.ucr.edu/~stelo/hashfilter/>)

Pooling work area





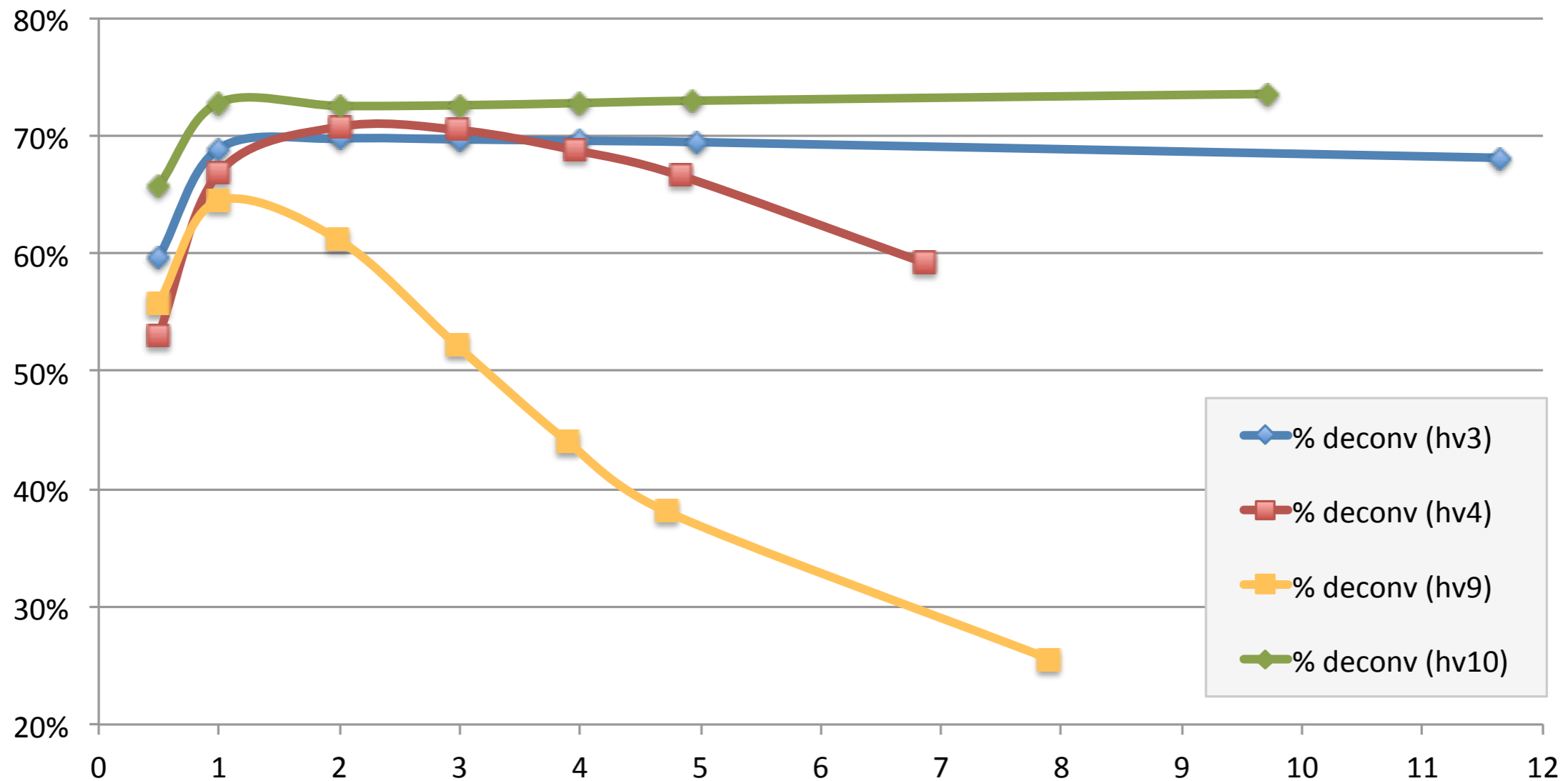
Pooling BAC barley clones



Sequencing BACs (Hv10)

- After demultiplexing
 - average of ~12.6M reads per pool
 - an average length of ~92 bases
- After trimming, adapter and *E.coli* removal
 - average of ~9.7M reads per pool
 - average length of ~90 bases
- Given that average BAC length in Hv10 is ~128kb, the average sequencing depth (before deconvolution) is ~500x

When “less is more”: slicing the data



- Too *few* reads per pool do not allow for decoding to work
- Too *many* reads per pool negatively affect the decoding due to sequencing errors

HvI0 decoding results

- HashFilter decoded 84.6% of the reads which translated into an average BAC sequencing depth of ~499x
[*time: ~7h, memory: 36.5 Gb*]
- Accuracy: ~21% of the BAC “signatures” in HvI0 were not used, HashFilter was not “aware” of it; only 0.043% of the reads were assigned to unused BAC signatures

BAC assembly

- Velvet assembled individual BACs, for ten different choices of the hash length parameter [Zerbino *et al.*, *Genome Res.* 2008]
- Recorded the statistics for the assembly that achieved the largest N50 (does not guarantee the 'best' overall assembly)
- [N50: the minimum length of all contigs/scaffolds that together account for at least 50% of the target]

BAC assembly statistics

- Hv10
 - N50 42,819 bp (36%)
 - largest contig 54,122 bp (45%)
 - sum of all contig sizes 147,639 bp (122%)
Average statistics over 1,053 BAC assemblies
- 3,237 BACs in Hv3-Hv10 were expected to contain known genes
- 2,877 (~89%) BAC assemblies contained the expected genes (with high coverage)

Comparing assemblies of one BAC

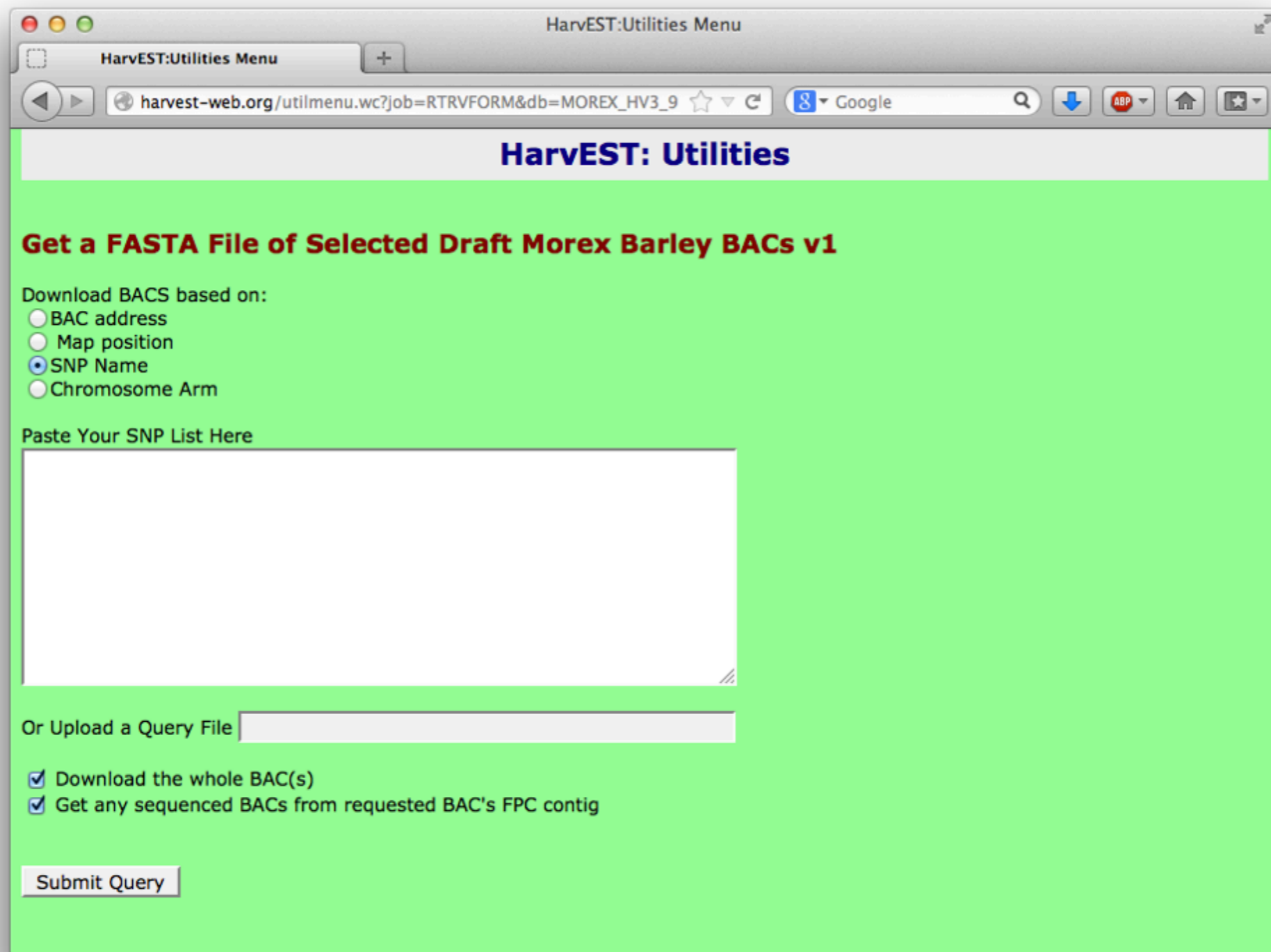
- BAC “0152010” has been sequenced
 - twice using combinatorial pooling (Hv3 and Hv9)
 - once as an individual BAC using Illumina
 - once using Sanger sequencing by JGI
- Using Sanger assembly as the “ground truth”, how the other three assemblies compare?

Comparing assemblies of one BAC

<i>assembly</i>	<i>Hv3</i>	<i>Hv9</i>	<i>single</i>
sequencing depth	~600x	~300x	~9,000x
# contigs	92	34	64
total length	146,889	124,772	240,997
largest contig	35,632	29,065	36,496
N50	20,290	16,959	34,200
# mis-assemblies	9	0	17
Genome fraction	89.6%	87.2%	73.4%

- Comparison produced with QUAST [<http://sourceforge.net/projects/quast/>]
- True BAC length is 131,747 bp

BAC assemblies: harvest-web.org



The screenshot shows a web browser window titled "HarvEST:Utilities Menu". The address bar displays the URL "harvest-web.org/utilmenu.wc?job=RTRVFORM&db=MOREX_HV3_9". The page has a light green background and a header bar with the text "HarvEST: Utilities" in blue. Below the header, the main heading is "Get a FASTA File of Selected Draft Morex Barley BACs v1". Underneath, there is a section "Download BACS based on:" with four radio button options: "BAC address", "Map position", "SNP Name" (which is selected), and "Chromosome Arm". Below these options is a text input field labeled "Paste Your SNP List Here". Further down, there is a section "Or Upload a Query File" with a file upload button. At the bottom, there are two checked checkboxes: "Download the whole BAC(s)" and "Get any sequenced BACs from requested BAC's FPC contig". A "Submit Query" button is located at the very bottom of the form.

HarvEST:Utilities Menu

Harvest-web.org/utilmenu.wc?job=RTRVFORM&db=MOREX_HV3_9

Google

HarvEST: Utilities

Get a FASTA File of Selected Draft Morex Barley BACs v1

Download BACS based on:

- ☐ BAC address
- ☐ Map position
- ☒ SNP Name
- ☐ Chromosome Arm

Paste Your SNP List Here

Or Upload a Query File

- ☒ Download the whole BAC(s)
- ☒ Get any sequenced BACs from requested BAC's FPC contig

Submit Query

For the latest version of BAC assemblies contact the presenters

Final remarks (1/2)

- BAC-by-BAC sequencing/assembly might be necessary for large, highly repetitive genomes
- BAC-by-BAC sequencing on NGS hinges on the ability of multiplexing hundreds of samples; DNA barcoding does not readily scale
- Combinatorial pooling is cost-effective and practical alternative to exhaustive DNA barcoding (both can be combined)

Final remarks (2/2)

- Experimental results confirm that the deconvolution process is very accurate
- Resulting BAC assemblies have high quality
- If the MTP set is given, cost is \$10-25/BAC (pooling, DNA preps, sequencing, informatics)
- Barley BACs and software are available
- Manuscripts
 - *PLoS Comp Biology*, 2013
 - *Proc. Workshop on Algorithms in Bioinformatics*, 2013

Acknowledgements

Botany and Plant Sciences, UC Riverside

Timothy Close (supervision, BACs, libraries, sequencing)

Steve Wanamaker (sys admin, read demux/cleaning)

Prasanna Bhat (Illumina OPA)

Yaqin Ma (sequencing library prep)

Josh Resnik (BAC pooling)

Computer Science, UC Riverside

Stefano Lonardi (supervision, deconvolution, assemblies)

Gianfranco Ciardo (deconvolution)

Denisa Duma (rice synthetic data, deconvolution)

Matthew Alpert (assemblies)

Burair Alsaihati (deconvolution)

Yonghui Wu (Illumina OPA deconvolution)

Serdar Bozdog (compartmentalized physical map)

Computer Science, University of Torino

Francesca Cordero (HashFilter)

Marco Beccuti (HashFilter)

Plant Sciences, UC Davis

Ming-Cheng Luo (*fingerprinting*)

Department of Energy, Joint Genome Institute

Jeremy Schmutz (Sanger sequencing)

Jane Grimwood (Sanger sequencing)



2009-65300-05645

DBI-1062301

**Please fill out the survey evaluation.
You will be contacted via email.**

Today's Presentation Available
<http://www.extension.org/pages/67926>

Sign up for PBG News
<http://pbgworks.org>

Sign up for Future Webinars and View Archive
<http://www.extension.org/pages/60426>

