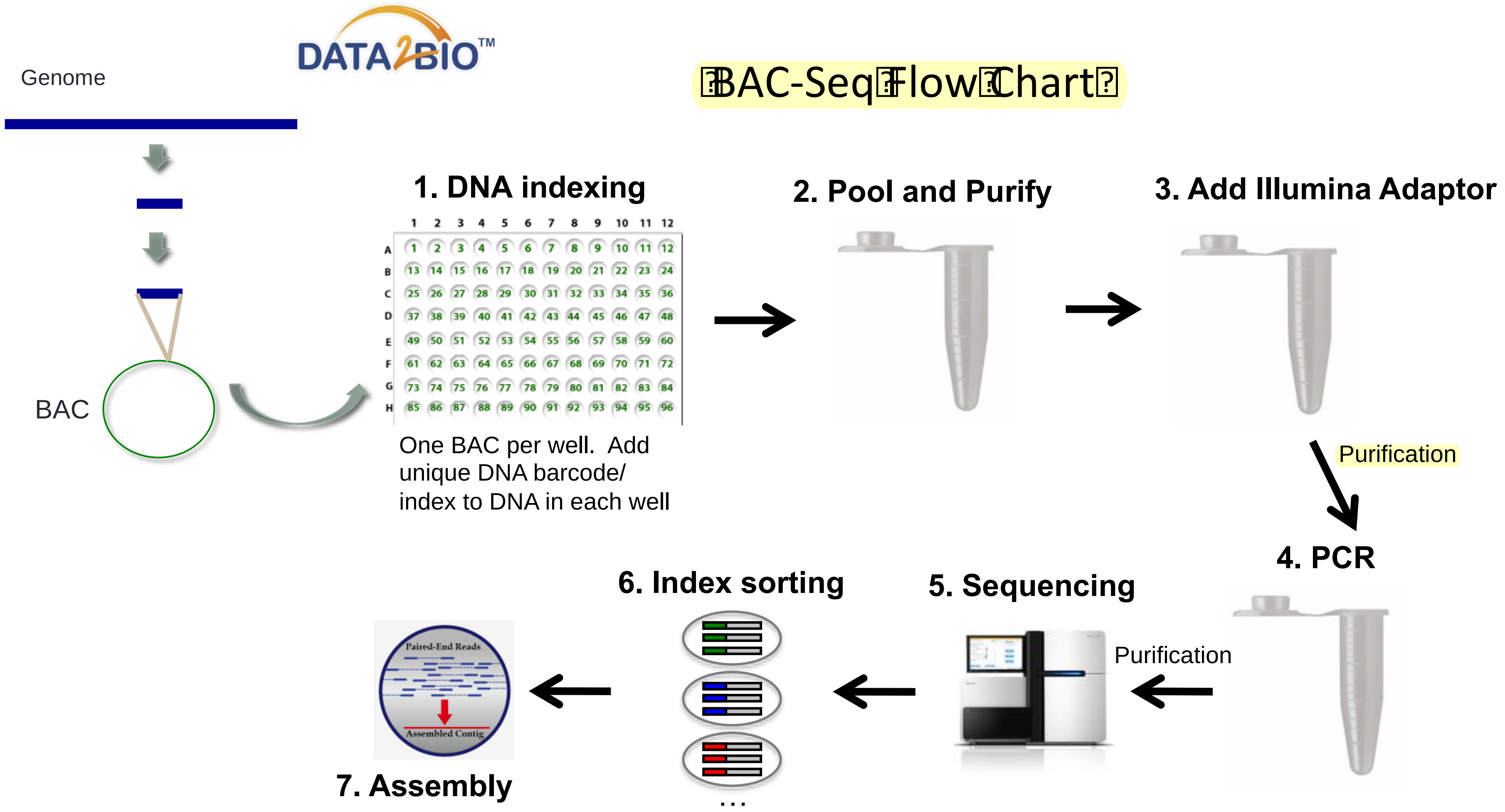
BACSEQ PROJECT UPDATE 02/07/14

Hung-Ying Lin

Introduction



	1	2	3	4	5	6	7	8	9	10	11	12
A	1	2	3	4	5	6	7	8	9	10	11	12
B	13	14	15	16	17	18	19	20	21	22	23	24
C	25	26	27	28	29	30	31	32	33	34	35	36
D	37	38	39	40	41	42	43	44	45	46	47	48
E	49	50	51	52	53	54	55	56	57	58	59	60
F	61	62	63	64	65	66	67	68	69	70	71	72
G	73	74	75	76	77	78	79	80	81	82	83	84
H	85	86	87	88	89	90	91	92	93	94	95	96

One BAC per well. Add unique DNA barcode/ index to DNA in each well

4 Res: BanII, Bsp1286I, NspI, NlaIII

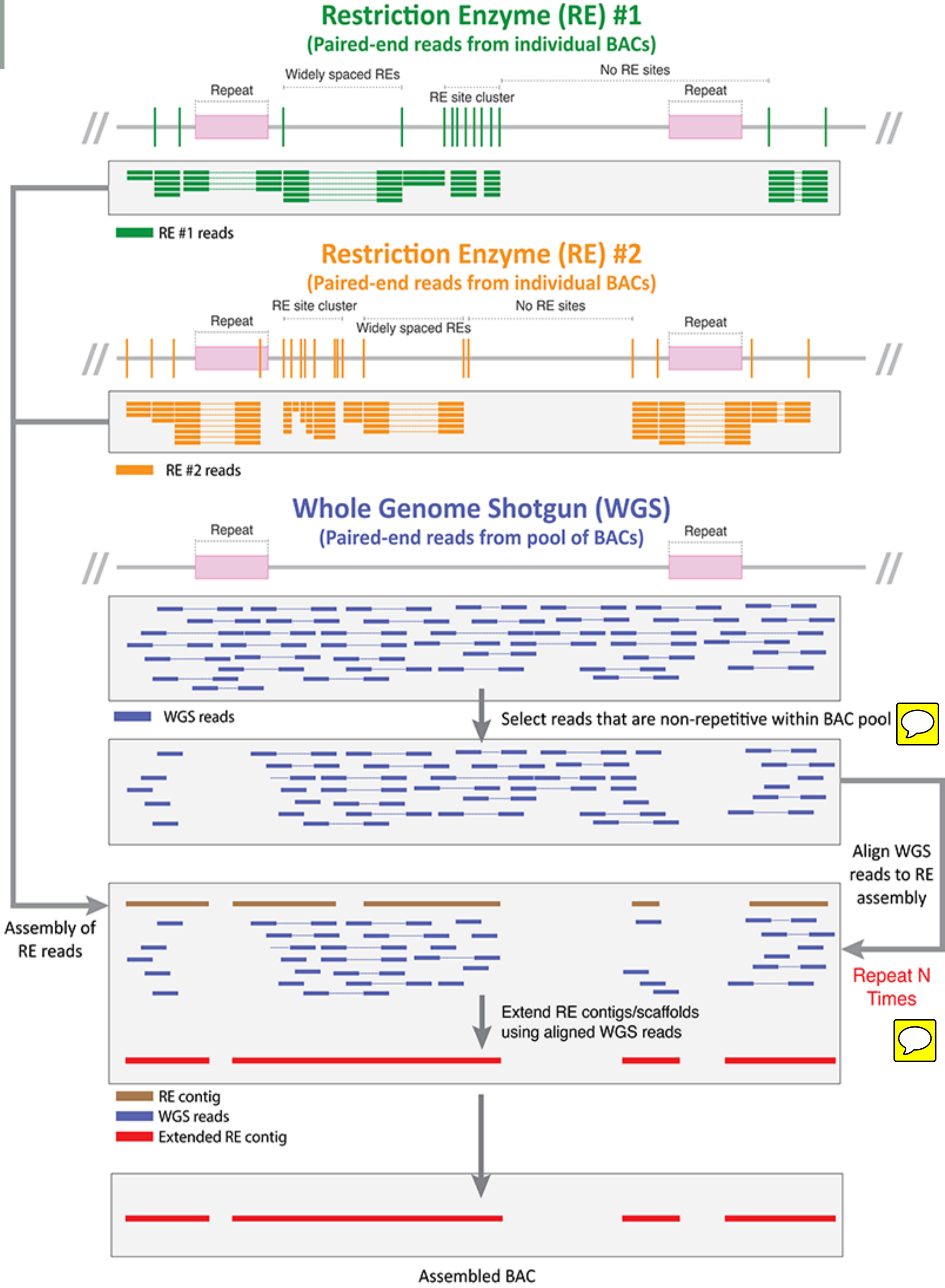
Data:

1. Reference: 12 B73 BACs

2. Restriction Enzyme reads about 10M reads (real data)

3. Real Whole Genome Shotgun reads
   - Miseq 2 * 250 bp
   - Enzyme shearing: Chimeric reads (2.6M; depth~400)
   - Physical shearing: Fragment size only 375bp (1.3M; depth~200)

4. Simulated Whole genome data
   - Miseq 2 * 250 bp
   - Fragment size: 750bp (1.4M; depth~ 200)
   - Fragment size: 1,500bp (0.7M; depth~ 100)



Restriction Enzyme (RE) #1
(Paired-end reads from individual BACs)

Repeat

Widely spaced REs

RE site cluster

No RE sites

Repeat

🟩 RE #1 reads

Restriction Enzyme (RE) #2
(Paired-end reads from individual BACs)

Repeat

RE site cluster

Widely spaced REs

No RE sites

Repeat

🟧 RE #2 reads

Whole Genome Shotgun (WGS)
(Paired-end reads from pool of BACs)

Repeat

Repeat

🟦 WGS reads

Select reads that are non-repetitive within BAC pool

Align WGS reads to RE assembly

Repeat N Times

Assembly of RE reads

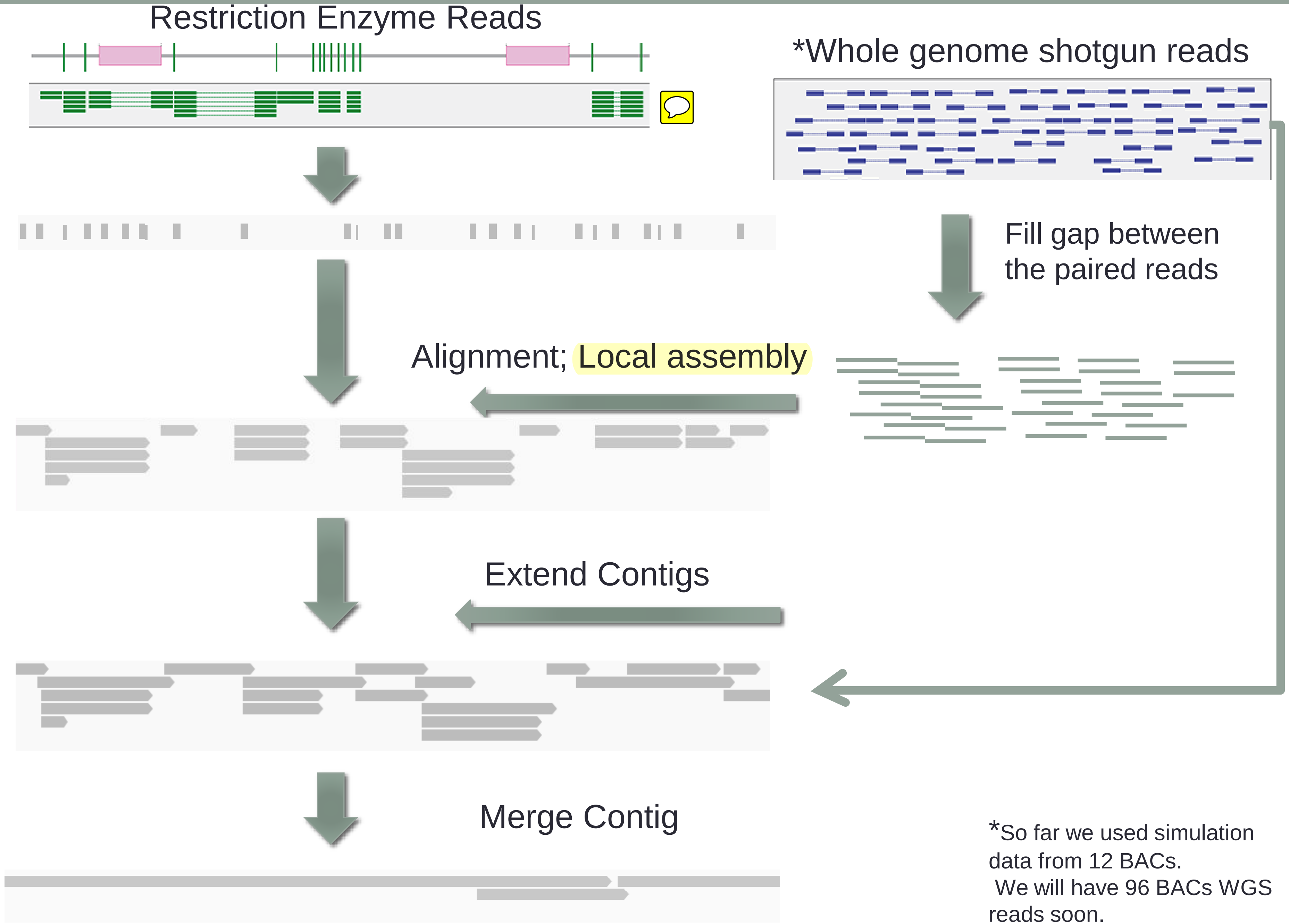
Extend RE contigs/scaffolds using aligned WGS reads

🟫 RE contig

🟦 WGS reads

🟥 Extended RE contig

Assembled BAC



Restriction Enzyme Reads

*Whole genome shotgun reads

Fill gap between the paired reads

Alignment; Local assembly

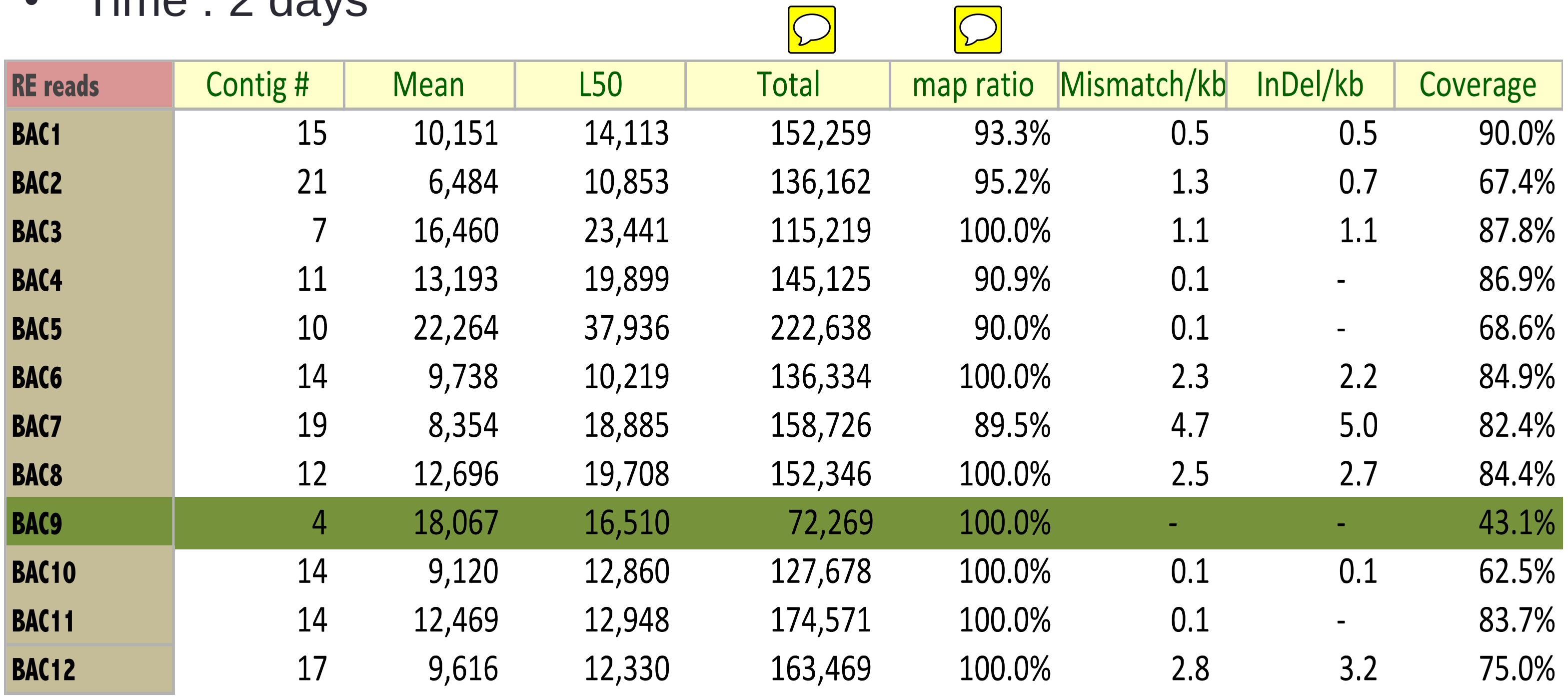
Extend Contigs

Merge Contig

*So far we used simulation data from 12 BACs.
We will have 96 BACs WGS reads soon.

BACseq Pipeline Updated Result

- Data: Fragment 750bp 200fold; Fragment 1,500bp 100fold (300fold in total)
- Cores: 12
- Time : 2 days



RE reads	Contig #	Mean	L50	Total	map ratio	Mismatch/kb	InDel/kb	Coverage
BAC1	15	10,151	14,113	152,259	93.3%	0.5	0.5	90.0%
BAC2	21	6,484	10,853	136,162	95.2%	1.3	0.7	67.4%
BAC3	7	16,460	23,441	115,219	100.0%	1.1	1.1	87.8%
BAC4	11	13,193	19,899	145,125	90.9%	0.1	-	86.9%
BAC5	10	22,264	37,936	222,638	90.0%	0.1	-	68.6%
BAC6	14	9,738	10,219	136,334	100.0%	2.3	2.2	84.9%
BAC7	19	8,354	18,885	158,726	89.5%	4.7	5.0	82.4%
BAC8	12	12,696	19,708	152,346	100.0%	2.5	2.7	84.4%
BAC9	4	18,067	16,510	72,269	100.0%	-	-	43.1% *
BAC10	14	9,120	12,860	127,678	100.0%	0.1	0.1	62.5%
BAC11	14	12,469	12,948	174,571	100.0%	0.1	-	83.7%
BAC12	17	9,616	12,330	163,469	100.0%	2.8	3.2	75.0%

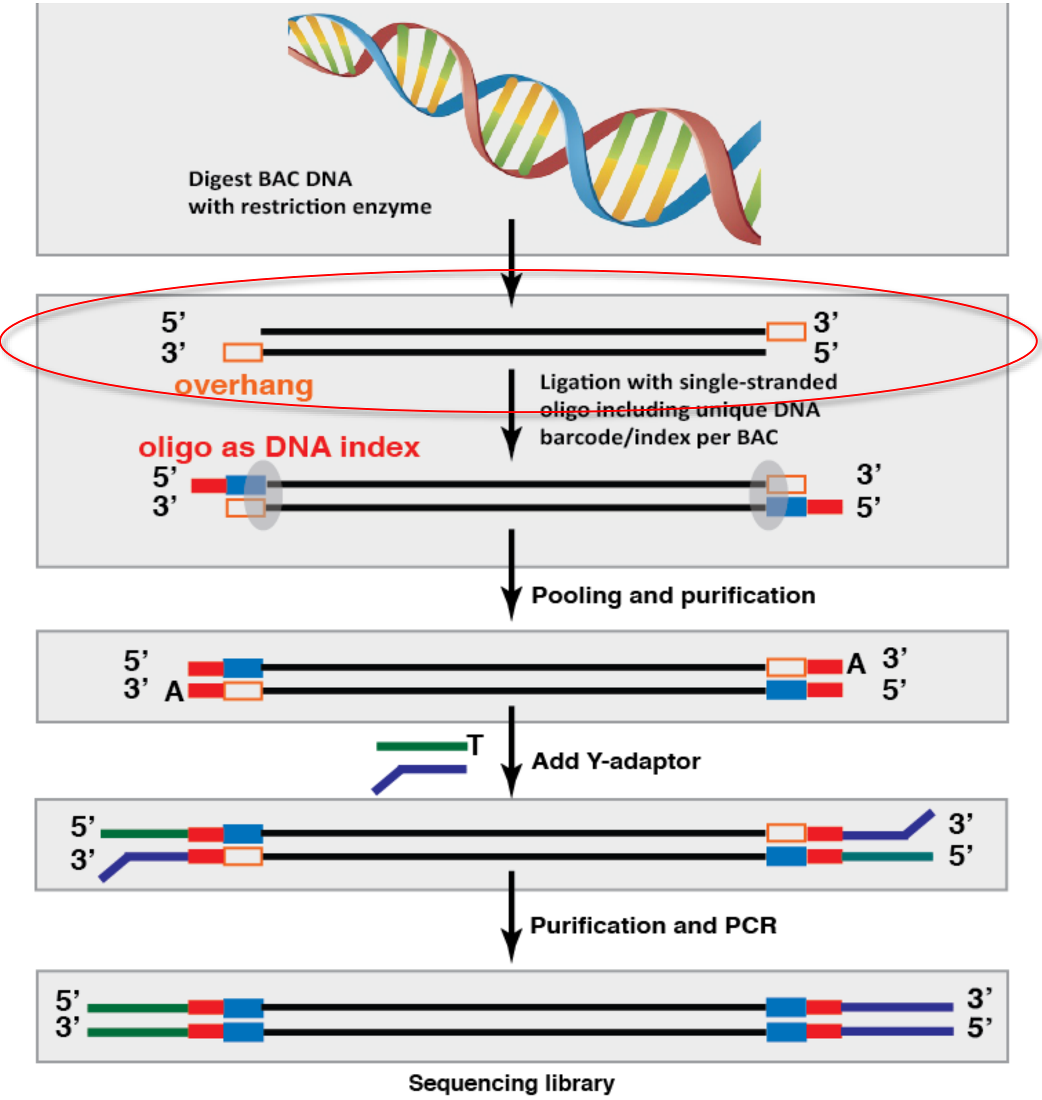
79.4% (not include BAC9)

* BAC9 restriction enzyme reads from other part of B73 genome sequence. That is not belong to all 12 BACs regions.

Limitations in BACseq Pipeline

- TASR
  - Accuracy
    - That will produce some chimeric contigs when (Contigs N50 > 10,000 bp)
  - Efficiency
    - For few BACs will takes a lot of time to finish target assembly
- Results are highly dependent on Restriction Enzyme Reads
  - Change new enzyme to avoid chimeric reads

Restriction Enzyme Part



Digest BAC DNA with restriction enzyme

Ligation with single-stranded oligo including unique DNA barcode/index per BAC

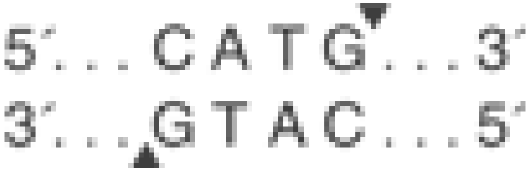
Pooling and purification

Add Y-adaptor

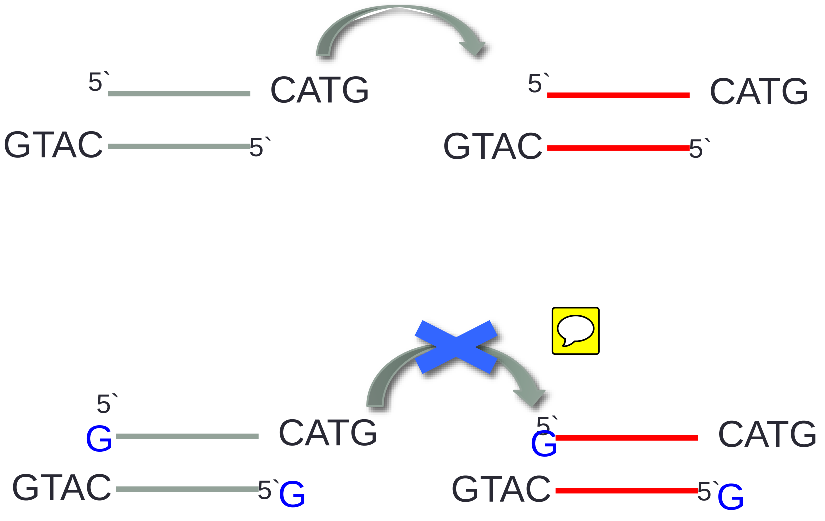
Purification and PCR

Sequencing library

- NlaIII (256bp)



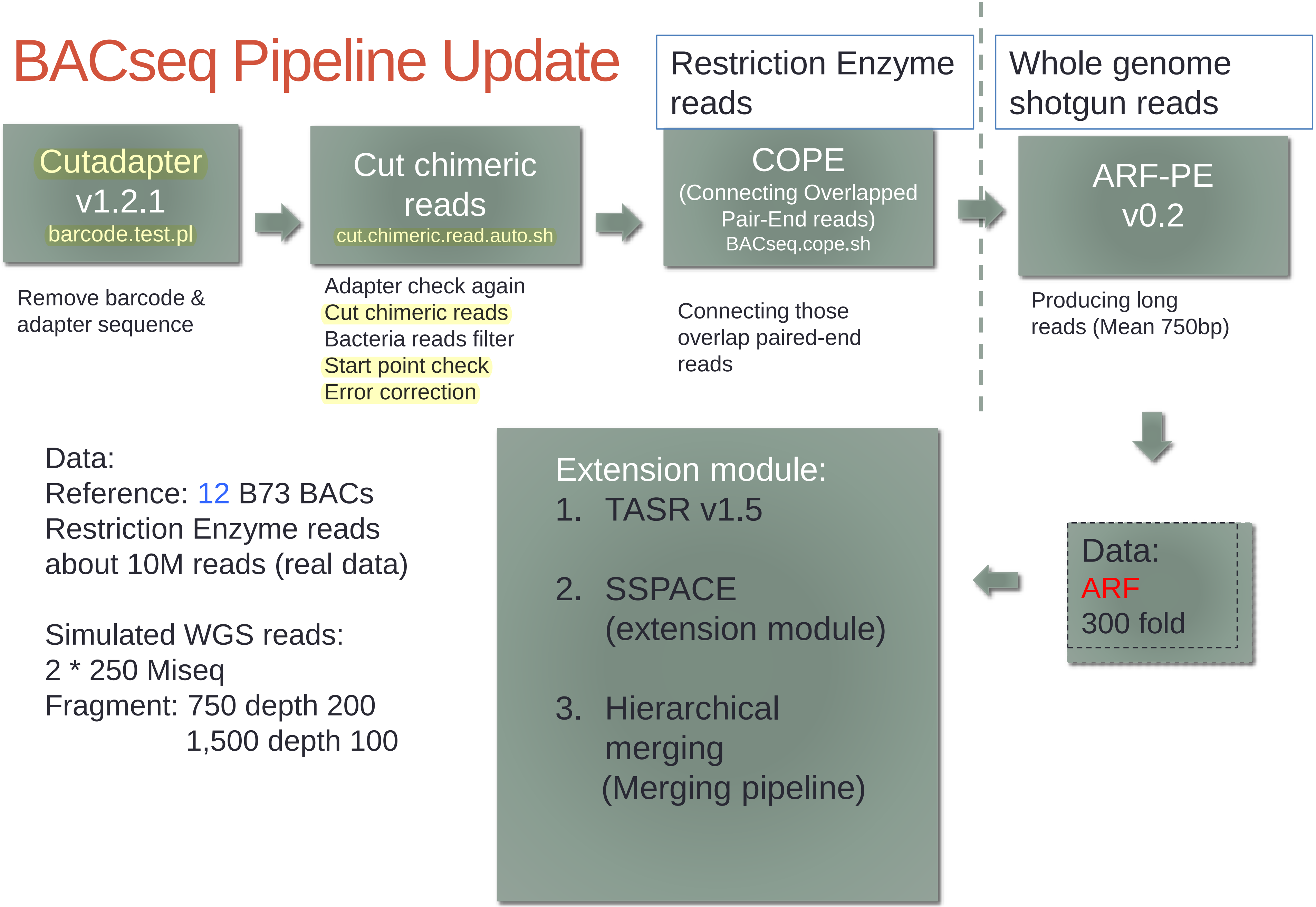
5'...CATG...3'
3'...GTAC...5'



Thank you so much !!

Following slides are other details

BACseq Pipeline Update



Restriction Enzyme reads

Whole genome shotgun reads

Cutadapter v1.2.1
barcode.test.pl

Remove barcode & adapter sequence

Cut chimeric reads
cut.chimeric.read.auto.sh

Adapter check again
Cut chimeric reads
Bacteria reads filter
Start point check
Error correction

COPE
(Connecting Overlapped Pair-End reads)
BACseq.cope.sh

Connecting those overlap paired-end reads

ARF-PE v0.2

Producing long reads (Mean 750bp)

Data:
ARF
300 fold

Extension module:

1. TASR v1.5

2. SSPACE
   (extension module)

3. Hierarchical merging
   (Merging pipeline)

Data:
Reference: 12 B73 BACs
Restriction Enzyme reads
about 10M reads (real data)

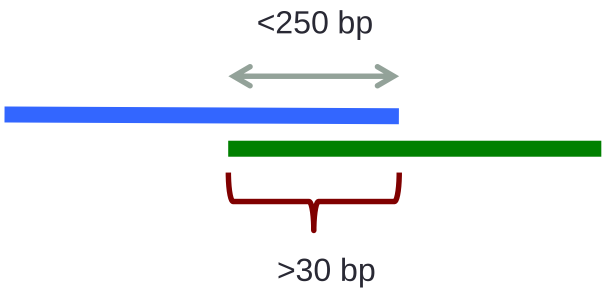
Simulated WGS reads:
2 * 250 Miseq
Fragment: 750 depth 200
1,500 depth 100

RE reads

- Connecting overlapped pair-end reads

(Restriction enzyme cut reads)

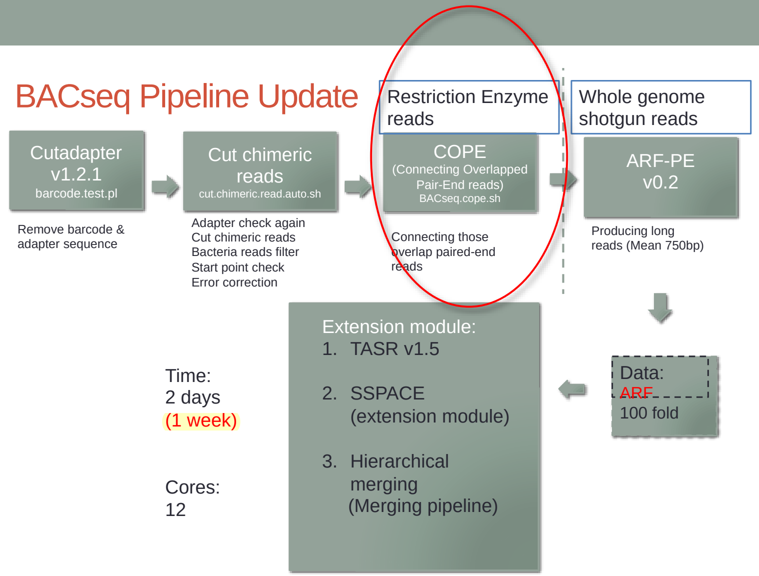
- COPE version v1.1.3



Identity > 95%

- Chimeric reads checking
  - Indexing long reads
  - Bowtie2 alignment

- Reads condense
  - In-house script
  - Log10 scale



BACseq Pipeline Update

Restriction Enzyme reads

Whole genome shotgun reads

Cutadapter v1.2.1
barcode.test.pl

Remove barcode & adapter sequence

Cut chimeric reads
cut.chimeric.read.auto.sh

Adapter check again
Cut chimeric reads
Bacteria reads filter
Start point check
Error correction

COPE
(Connecting Overlapped Pair-End reads)
BACseq.cope.sh

Connecting those overlap paired-end reads

ARF-PE v0.2

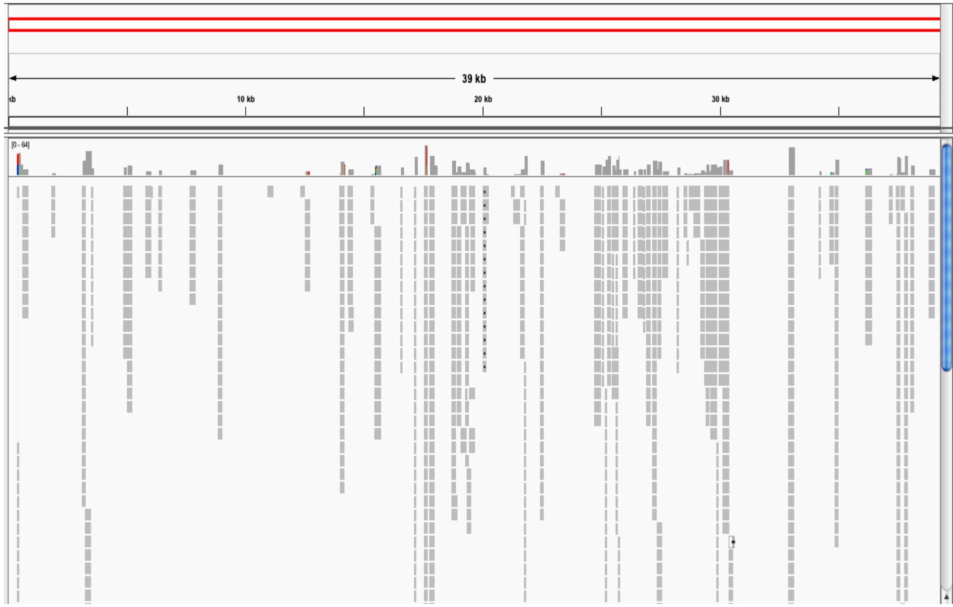
Producing long reads (Mean 750bp)

Data:
~~ARF~~
100 fold

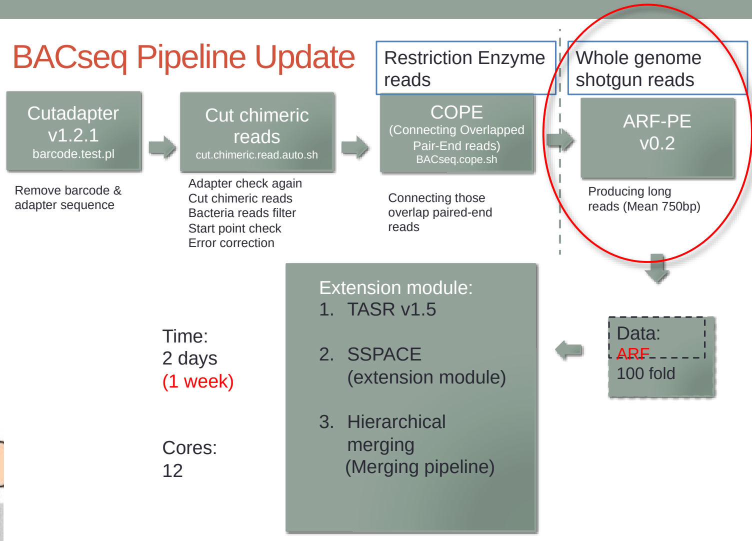
Extension module:
1. TASR v1.5
2. SSPACE (extension module)
3. Hierarchical merging (Merging pipeline)

Time:
2 days
(1 week)

Cores:
12



ARF-PE workflow



BACseq Pipeline Update

Restriction Enzyme reads

Whole genome shotgun reads

Cutadapter v1.2.1
barcode.test.pl

Remove barcode & adapter sequence

Cut chimeric reads
cut.chimeric.read.auto.sh

Adapter check again
Cut chimeric reads
Bacteria reads filter
Start point check
Error correction

COPE
(Connecting Overlapped Pair-End reads)
BACseq.cope.sh

Connecting those overlap paired-end reads

ARF-PE v0.2

Producing long reads (Mean 750bp)

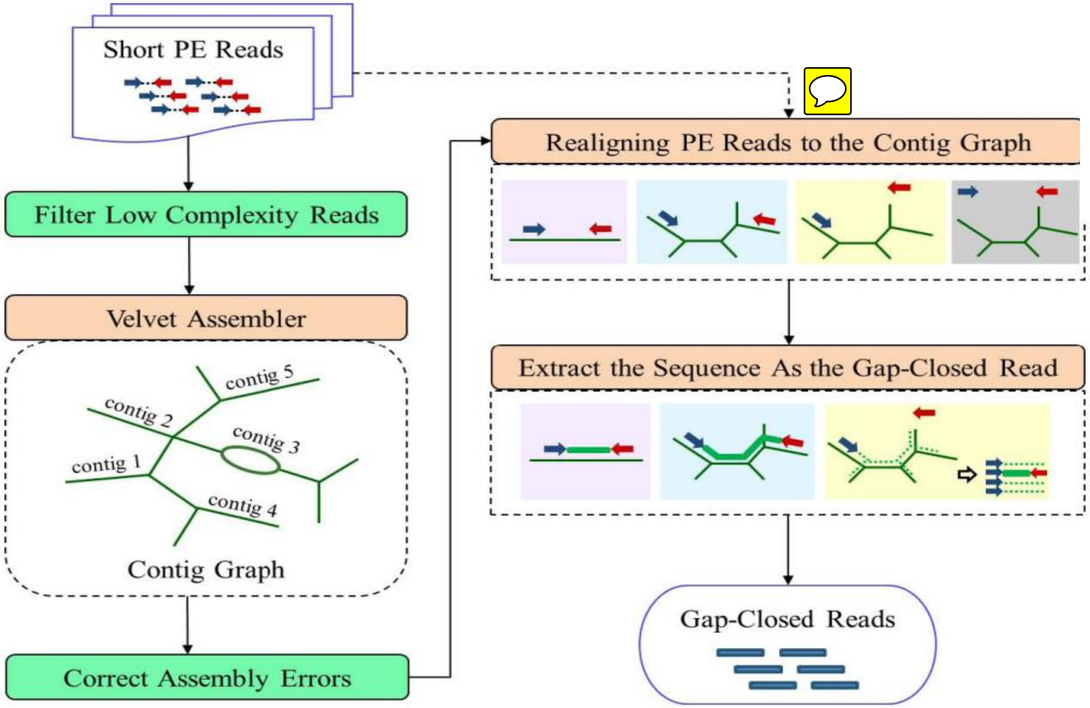
Data:
ARF
100 fold

Extension module:

1. TASR v1.5
2. SSPACE (extension module)
3. Hierarchical merging (Merging pipeline)

Time:
2 days
(1 week)

Cores:
12



- ARF-PE result

RE reads	Read #	Mean	L50	Total	map ratio	Mismatch/k	InDel/kb
BAC12	398,299	748	755	297,853,717	99.9%	0.0	0.0

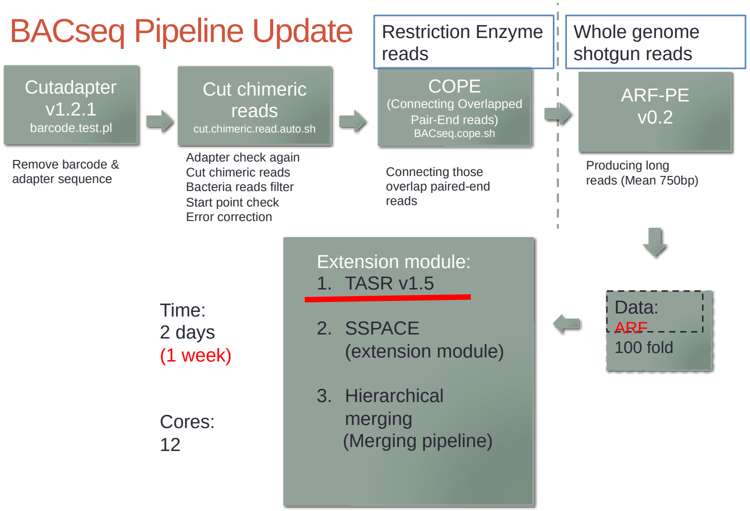
TASR(Target Assembly of Sequence Reads)

- TASR (version1.5)
- Input data:

Reads: Reads from ARF

Target sequence: RE reads contigs

- Overlap: 100 bp
  - Minimum number of reads needed to call a base: 1
  - Minimum base ratio used to accept a overhang: 0.9
- Iteration control:
  - Coverage:
    - If the coverage can not increase, this loop will be stop.



BACseq Pipeline Update

Restriction Enzyme reads

Whole genome shotgun reads

Cutadapter v1.2.1
barcode.test.pl

Remove barcode & adapter sequence

Cut chimeric reads
cut.chimeric.read.auto.sh

Adapter check again
Cut chimeric reads
Bacteria reads filter
Start point check
Error correction

COPE
(Connecting Overlapped Pair-End reads)
BACseq.cope.sh

Connecting those overlap paired-end reads

ARF-PE v0.2

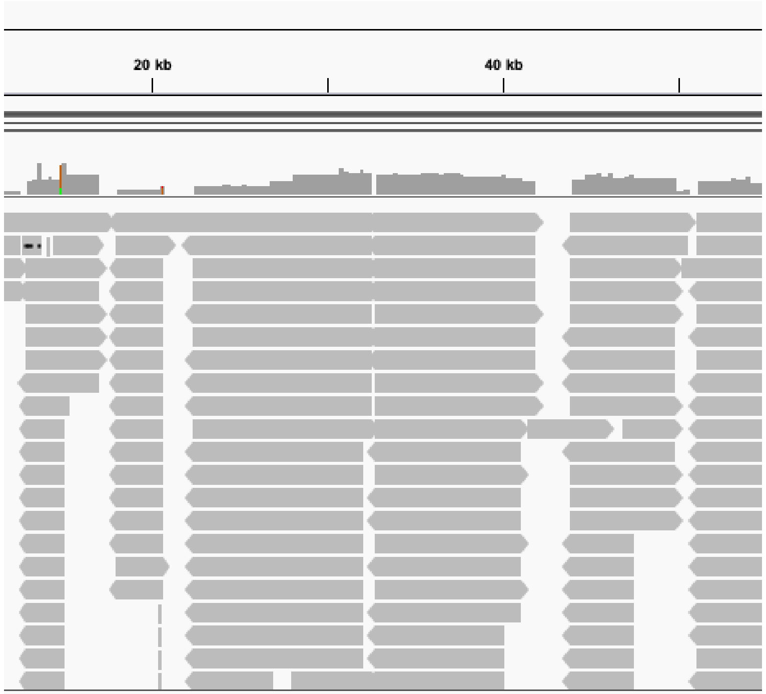
Producing long reads (Mean 750bp)

Data:
ARF
100 fold

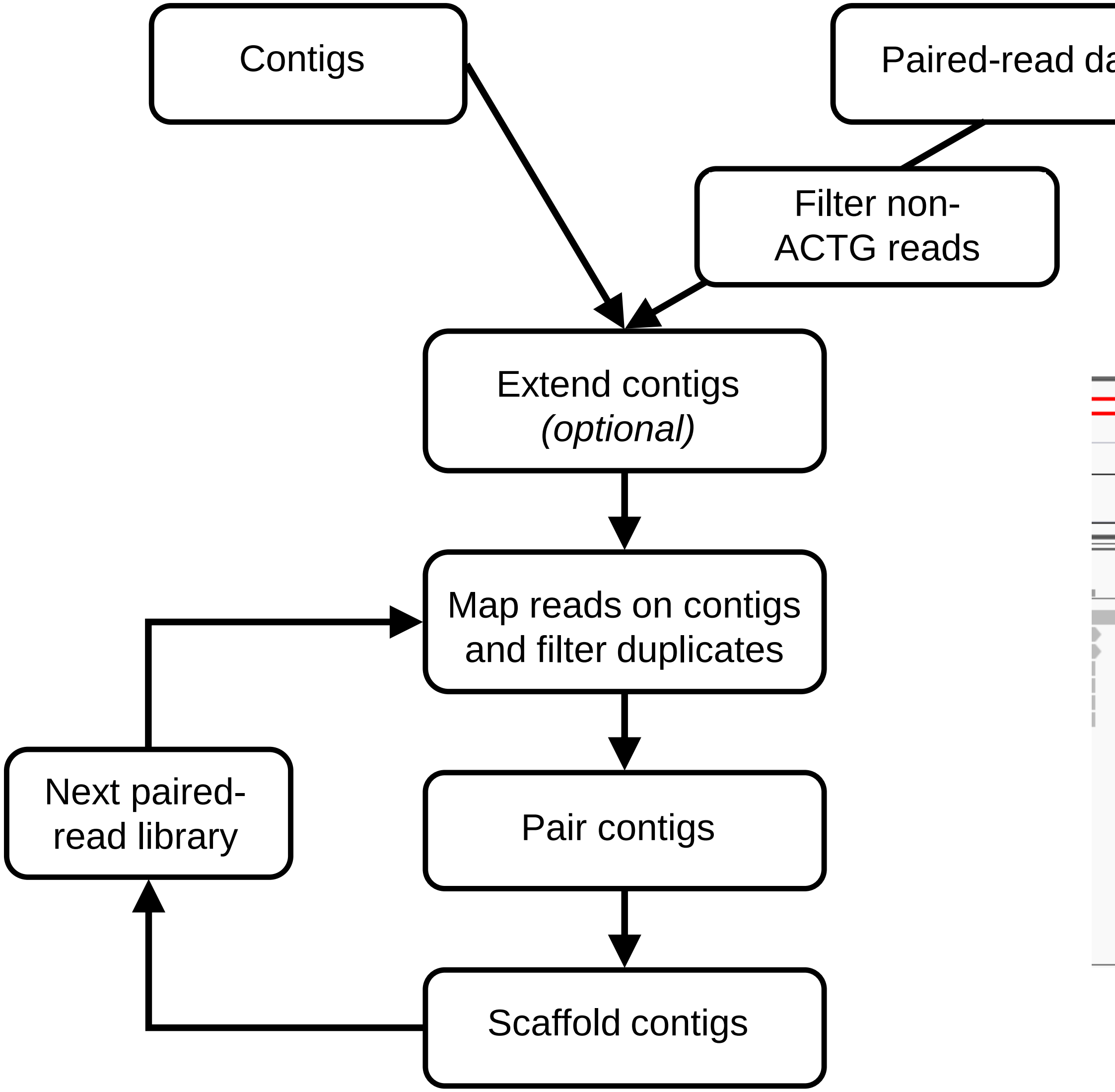
Extension module:
1. TASR v1.5
2. SSPACE (extension module)
3. Hierarchical merging (Merging pipeline)

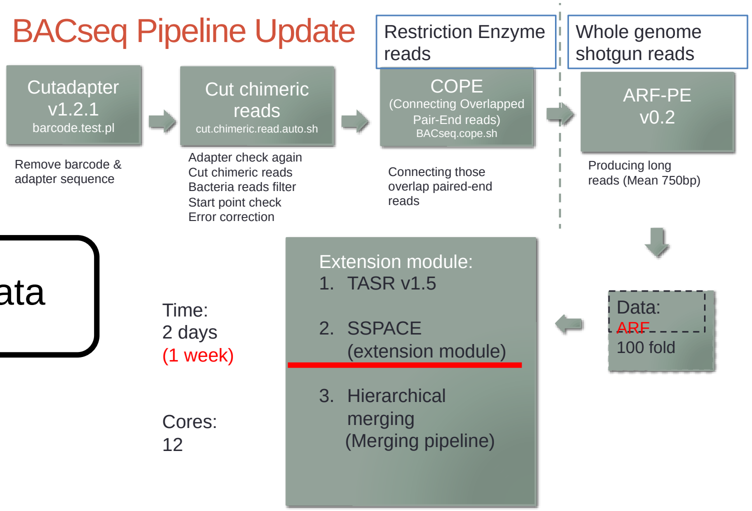
Time:
2 days
(1 week)

Cores:
12



SSPACE





BACseq Pipeline Update

Remove barcode & adapter sequence

Adapter check again
Cut chimeric reads
Bacteria reads filter
Start point check
Error correction

Connecting those overlap paired-end reads

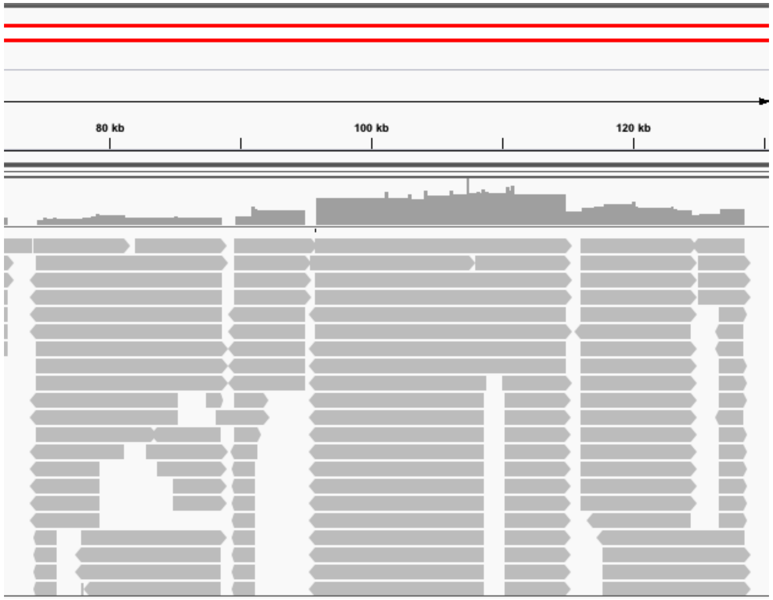
Producing long reads (Mean 750bp)

Time:
2 days
(1 week)

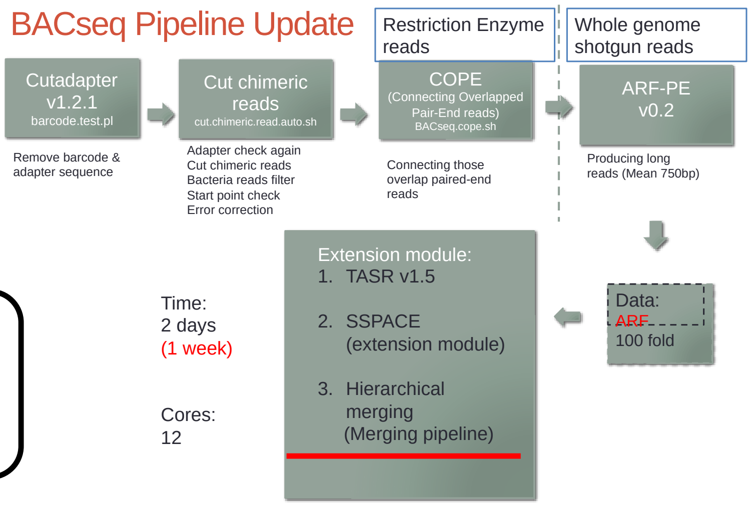
Cores:
12

Extension module:
1. TASR v1.5
2. SSPACE (extension module)
3. Hierarchical merging (Merging pipeline)

Data:
ARF
100 fold



Hierarchical Clustering



BACseq Pipeline Update

Restriction Enzyme reads

Whole genome shotgun reads

Cutadapter v1.2.1
barcode.test.pl

Remove barcode & adapter sequence

Cut chimeric reads
cut.chimeric.read.auto.sh

Adapter check again
Cut chimeric reads
Bacteria reads filter
Start point check
Error correction

COPE
(Connecting Overlapped Pair-End reads)
BACseq.cope.sh

Connecting those overlap paired-end reads

ARF-PE v0.2

Producing long reads (Mean 750bp)

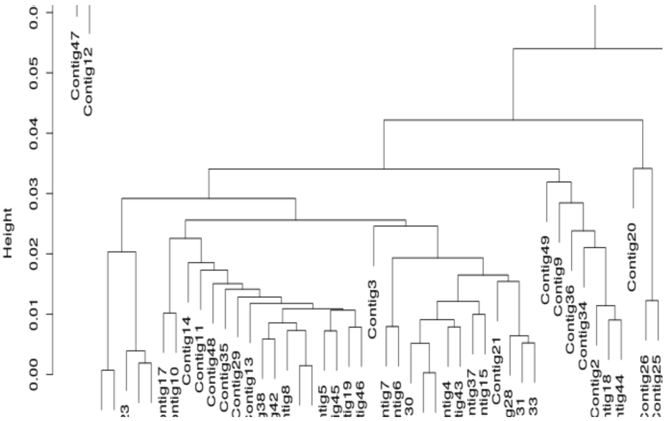
Data:
ARF
100 fold

Extension module:

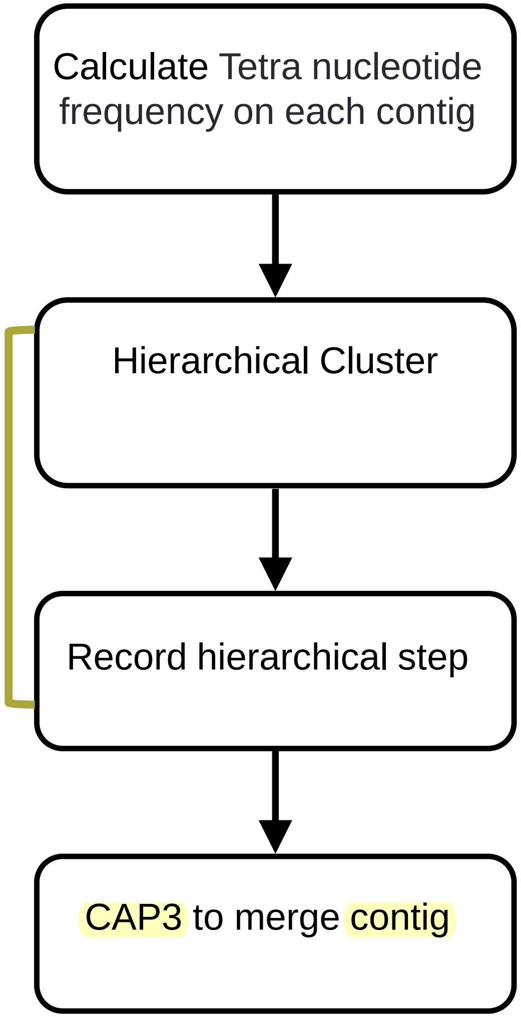
1. TASR v1.5
2. SSPACE (extension module)
3. Hierarchical merging (Merging pipeline)

Time:
2 days
(1 week)

Cores:
12



	X1	X2	X3
1	Contig.1	Contig.40	step.1
2	Contig.2	Contig.42	step.2
3	Contig.33	Contig.23	step.3
4	Contig.17	Contig.41	step.4
5	Contig.24	step.4	step.5
6	Contig.31	step.1	step.6
7	Contig.32	Contig.34	step.7
8	Contig.39	Contig.43	step.8
9	Contig.29	step.7	step.9
10	Contig.6	Contig.46	step.10



Calculate Tetra nucleotide frequency on each contig

Hierarchical Cluster

Record hierarchical step

CAP3 to merge contig