GENOME INSTITUTE

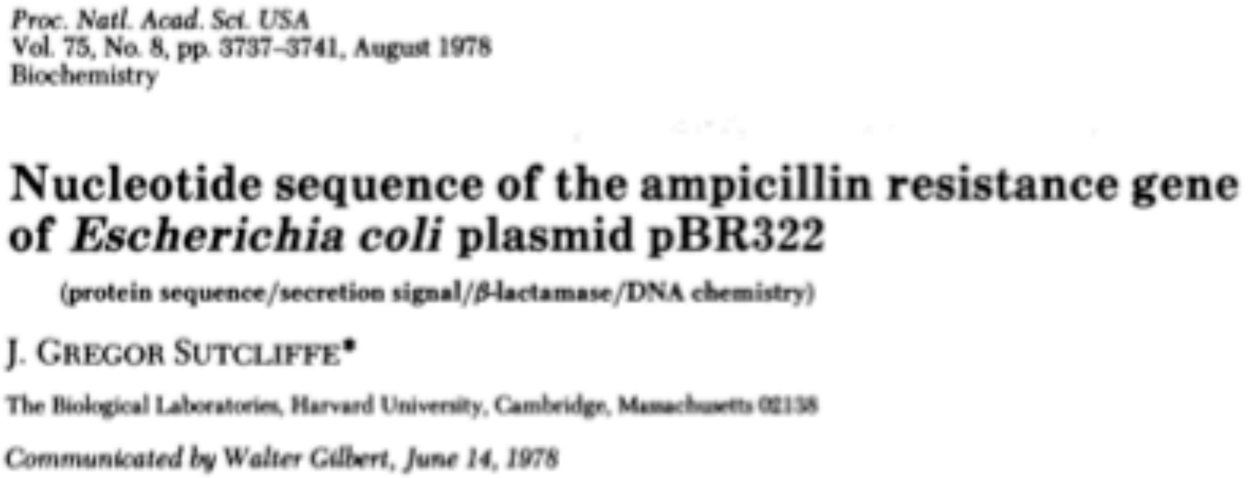
Towards the Perfect Genome Sequence

George Weinstock

Sequencing, Finishing, Analysis in the Future

Santa Fe, June 2012

Getting a sequence perfect - In the Beginning



Proc. Natl. Acad. Sci. USA
Vol. 75, No. 8, pp. 3737–3741, August 1978
Biochemistry

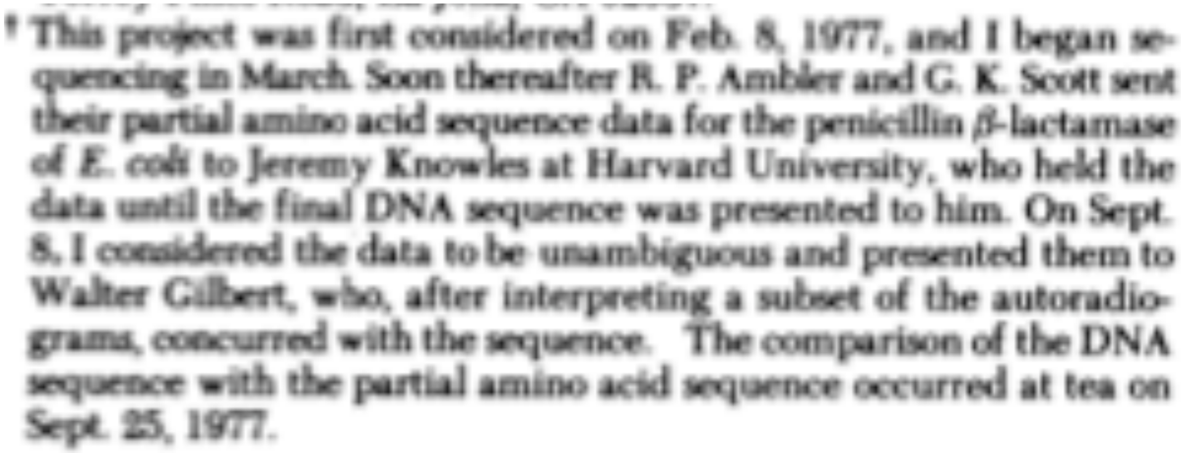
Nucleotide sequence of the ampicillin resistance gene of *Escherichia coli* plasmid pBR322

(protein sequence/secretion signal/β-lactamase/DNA chemistry)

J. GREGOR SUTCLIFFE*

The Biological Laboratories, Harvard University, Cambridge, Massachusetts 02138

Communicated by Walter Gilbert, June 14, 1978



* This project was first considered on Feb. 8, 1977, and I began sequencing in March. Soon thereafter R. P. Ambler and G. K. Scott sent their partial amino acid sequence data for the penicillin β-lactamase of *E. coli* to Jeremy Knowles at Harvard University, who held the data until the final DNA sequence was presented to him. On Sept. 8, I considered the data to be unambiguous and presented them to Walter Gilbert, who, after interpreting a subset of the autoradiograms, concurred with the sequence. The comparison of the DNA sequence with the partial amino acid sequence occurred at tea on Sept. 25, 1977.

Sequence BOTH strands!!

Finished genomes aren’t finished (!)

- “Finished” genomes have errors
- Multiple chromosomes, circular and linear; plasmids
  - Closure or not
- Definition of Finished was 10^{-4} to 10^{-5} base accuracy
  - Some regions are tricky; large regions may be error-free
- Mis-assemblies present; can be difficult to detect
  - Read pairs one way to show inconsistencies
- Correct placement of repeats a challenge
  - rRNA gene clusters
  - Mobile elements
  - Paralogs and other sequence families
  - Tandem and dispersed repeats each pose their own challenges

Treponema pallidum

The complete genome sequence of *Treponema pallidum* was determined and shown to be 1,138,006 base pairs containing 1041 predicted coding sequences (open reading frames). Systems for DNA replication, transcription, translation, and repair are intact, but catabolic and biosynthetic activities are minimized. The number of identifiable transporters is small, and no phosphoenolpyruvate:phosphotransferase carbohydrate transporters were found. Potential virulence factors include a family of 12 potential membrane proteins and several putative hemolysins. Comparison of the *T. pallidum* genome sequence with that of another pathogenic spirochete, *Borrelia burgdorferi*, the agent of Lyme disease, identified unique and common genes and substantiates the considerable diversity observed among pathogenic spirochetes.

Genome analysis.

C. M. Fraser, O. White, G. G. Sutton, R. Dodson, M. Gwinn, E. K. Hickey, R. Clayton, K. A. Ketchum, S. Salzberg, J. Peterson, H. Khalak, D. Richardson, T. Utterback, L. McDonald, P. Artiach, C. Bowman, M. D. Cotton, C. Fujii, S. Garland, B. Hatch, K. Horst, K. Roberts, M. Sandusky, J. Weidman, H. O. Smith, and J. C. Venter are with The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA. S. J. Norris, G. M. Weinstock, E. Sodergren, J. M. Hardham, M. P. McLeod, J. K. Howell, and M. Chidambaram are at the University of Texas Health Science Center, Departments of Microbiology and Molecular Genetics, Pathology and Laboratory Medicine, and the Center for the Study of Emerging and Reemerging Pathogens, Post Office Box 20708, Houston, TX 77225, USA.

*To whom correspondence should be addressed. E-mail: tpdb@tigr.org

1998

Treponema pallidum Nichols sequence (1.1 Mb) Since 1998 many errors found

- TP0126: *tprK*-like gene (1.3kb) that integrates/excises – not in original sequence.
  - Complex region: donor sites for *tprK*, at least several new genes
  - Undergoes sequence variation during growth
- TP0433-434: addition of 60bp in the *arp* gene.
  - 7 tandem repetitions in reference but correct number is 14 – some collapsed in original assembly.
- IGR TP0135-136: two populations of the Nichols strain – with and without insertion of 64 bp between these genes.
- Sequencing errors in 206 ORFs
  - 396 substitutions, 13 insertions, 9 deletions
- Still working on it after >15 years!! (David Šmajs et al)

Finishing approaches: Sanger

- Sanger sequencing – slow and expensive
- Sequencing clones in a tiling path – laborious
- Finishing shotgun clones – need templates etc.
- Kitchen sink approach
    - Manual joining of missed overlaps
    - Targeted PCR-sequencing to fill gaps
    - Very small insert libraries of poorly clonable or gnarly sequences
    - Alternative nucleotides to get through secondary structure
    - ETC!!
- Why can’t we do better?
    - 17 years of bacterial genome sequencing
    - Genome assembly software first developed for bacteria
    - Little/slow progress?

Along comes Next Generation Sequencing

- Cheaper, faster than Sanger
- Less manual work – highly parallel
- But
  - More errors
  - Short read lengths => challenging for repeats
  - Higher coverage required => polymorphism in culture apparent

The Challenge:

Can one finish a genome without finishing*, only NGS?

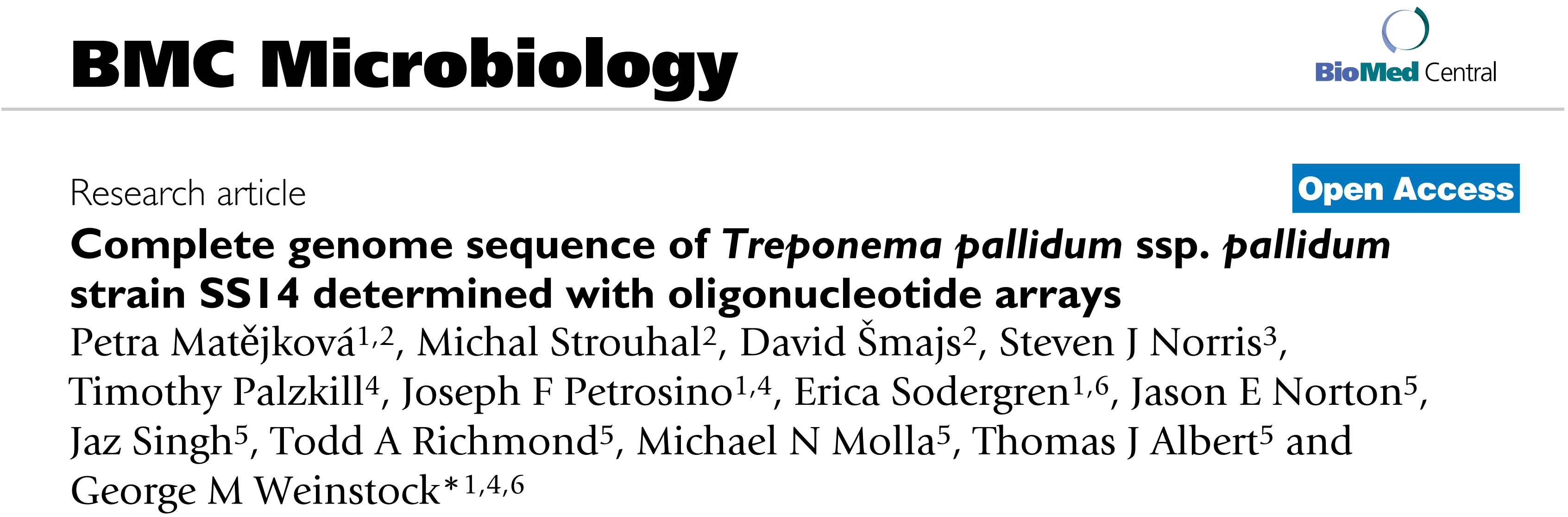
*Sanger finishing

Is there hope that this can work?

- “Upgrade” finished* genomes with new platforms
  - c2006: Advent of NGS: 454 GS20 (100bp), Illumina (36 bp)
  - *Treponema pallidum*, *Staphylococcus aureus*, *Escherichia coli*, *Francisella tularensis* tests
  - Assemble deep shotgun (Newbler, Velvet)
  - Compare to finished genomes with cross_match
    - When disagreement, majority rules
- ***Produces high quality (perfect?) base sequence***
- Mis-assemblies could still be an issue

* Finished with Sanger

Sequence of a second *T. pallidum* strain (SS14)
Use independent method (Comparative Genome Sequencing)



BMC Microbiology

BioMed Central

Research article

Open Access

Complete genome sequence of *Treponema pallidum* ssp. *pallidum* strain SS14 determined with oligonucleotide arrays

Petra Matějková[1,2], Michal Strouhal[2], David Šmajs[2], Steven J Norris[3], Timothy Palzkill[4], Joseph F Petrosino[1,4], Erica Sodergren[1,6], Jason E Norton[5], Jaz Singh[5], Todd A Richmond[5], Michael N Molla[5], Thomas J Albert[5] and George M Weinstock*[1,4,6]

Treponema pallidum type strains

subsp. *pallidum*

Baltimore
🟣 ⚫ DAL-1
Grady
🟣 Mexico A
MN-3
🟣 ⚫ Nichols
Philadelphia 1
Philadelphia 2
🟠 🟣 ⚫ Street strain 14

subsp. *pertenue*

CDC-1
🟣 ⚫ CDC-2
🟣 ⚫ Gauthier
🟠 🟣 ⚫ Samoa D
Samoa F

subsp. *endemicum*

🔴 🟣 ⚫ Bosnia A
Iraq B

PSGS

unclassified simian isolate

🟣 ⚫ Fribourg-Blanc

PSGS

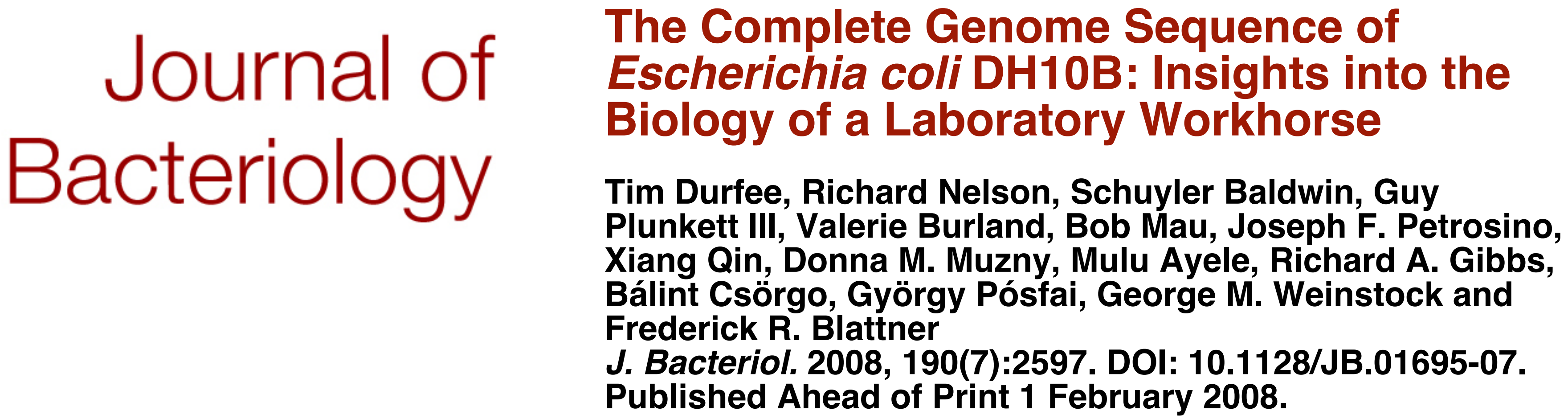
Treponema paraluiscuniculi

🟠 🟣 ⚫ Cuniculi A

⚫	**454 data**
🟣	**Solexa/Illumina data**
🟠	**CGS data**
🔴	**SOLiD data**

PSGS	Pooled segment genomic sequencing
CGS	Comparative genomic sequencing

What about mis-assemblies?



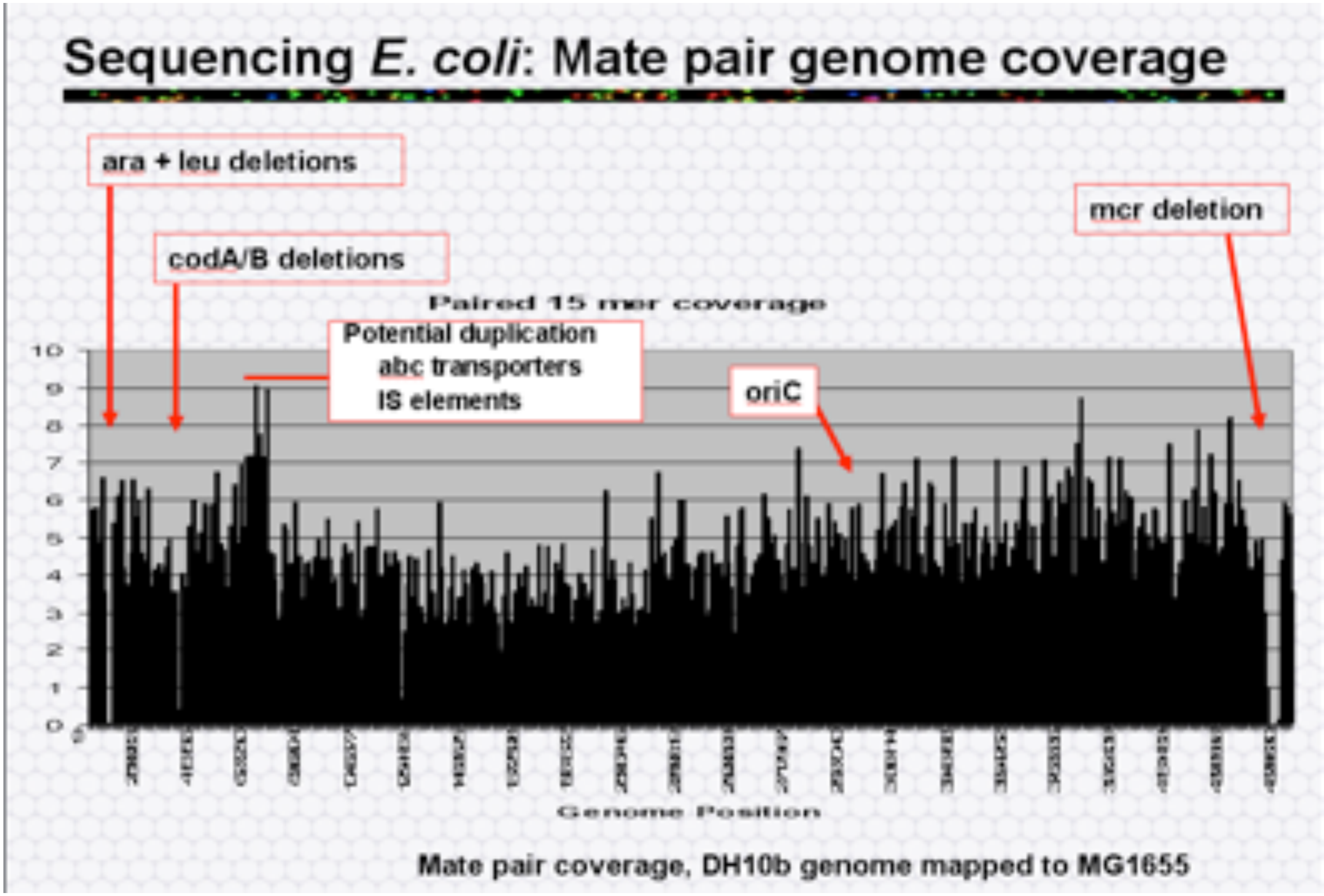
Journal of Bacteriology

The Complete Genome Sequence of *Escherichia coli* DH10B: Insights into the Biology of a Laboratory Workhorse

Tim Durfee, Richard Nelson, Schuyler Baldwin, Guy Plunkett III, Valerie Burland, Bob Mau, Joseph F. Petrosino, Xiang Qin, Donna M. Muzny, Mulu Ayele, Richard A. Gibbs, Bálint Csörgo, György Pósfai, George M. Weinstock and Frederick R. Blattner
***J. Bacteriol.* 2008, 190(7):2597. DOI: 10.1128/JB.01695-07.**
Published Ahead of Print 1 February 2008.

NGS (SOLiD) read pairs identify mis-assembly

E. coli DH10B: 113kb precise duplication
Collapsed in assembly of Sanger reads



Sequencing *E. coli*: Mate pair genome coverage

Mate pair coverage, DH10b genome mapped to MG1655

There is hope with NGS

What is the Perfect Genome?

- Topology: no gaps
- No mis-assemblies
- Correct bases

What is the Perfect Genome?

- Caveats: variants occur spontaneously in culture
  - Elements that insert/excise (e.g. E. coli e14 element)
  - Elements that invert (many examples of phase variation)
  - Sequence variation (e.g. antigenic variation)
  - Tandem duplications from recombination between repeats (rare)
- There may not be a single correct sequence for these
- NGS is deep sequencing: will pick these up
  - Challenge for assembly when polymorphisms present
- 8x Sanger would not routinely detect these variants
- So expect some intrinsic sequence ambiguity
- The perfect genome sequence should capture variations

Intrastrain heterogeneity seen at ~70x Solexa *T. pallidum* subspecies *pallidum* strain SS14

Genome	Position in ref seq	Base	Total coverage	Majority sequence	Corresponding coverage	Intrastrain heterogeneity	Gene
TPASS14	85401	G	43	insertion of G	39	G stretch	TP0077
TPASS14	135109	C or G	7				
TPASS14	135118	C or T	7				
TPASS14	135152	G or A	36	G	29	2 alternating bases	tprC
TPASS14	135155	C or T	28	T	26	2 alternating bases	tprC
TPASS14	135160	C or T	21	C	16	2 alternating bases	tprC
TPASS14	135231	G or A	25	A	19	2 alternating bases	tprC
TPASS14	135262	G or A	43	A	26	2 alternating bases	tprC
TPASS14	293812	T	54	insertion of T	53	T stretch	TP0277
TPASS14	673231	C or T	23	T	18	2 alternating bases	tprI
TPASS14	673236	G or T	20	T	16	2 alternating bases	tprI
TPASS14	673238	C or T	19	T	15	2 alternating bases	tprI
TPASS14	673248	C or T	33	C	21	2 alternating bases	tprI
TPASS14	673467	C or G	23	C	22	2 alternating bases	tprI
TPASS14	673489	C or T	13	T	13	2 alternating bases	tprI
TPASS14	673771	G or A	58	A	53	2 alternating bases	tprI
TPASS14	674430					ating bases	IGR
TPASS14	674911					ating bases	tprJ
TPASS14	674914	G or A	72	G	68	2 alternating bases	tprJ
TPASS14	675036	G or A	21	G	18	2 alternating bases	tprJ
TPASS14	675040	C or T	34	T	22	2 alternating bases	tprJ
TPASS14	831822	G	34	insertion of G	33	G stretch	IGR
TPASS14	870951	A	62	deletion of A	58	A stretch	TP0801
TPASS14	925246	C	46	insertion of C	44	C stretch	IGR
TPASS14	1063808	C or T	41	C	30	2 alternating bases	TP0979
TPASS14	1125302	G	46	insertion of G	44	G stretch	TP1029

TP0077 Stretch of G: 4/43 have an deletion of a G

tprI G/A polymorphism: 5/53 have G

Colour legend
2 alternating nucleotides
homopolymeric stretches

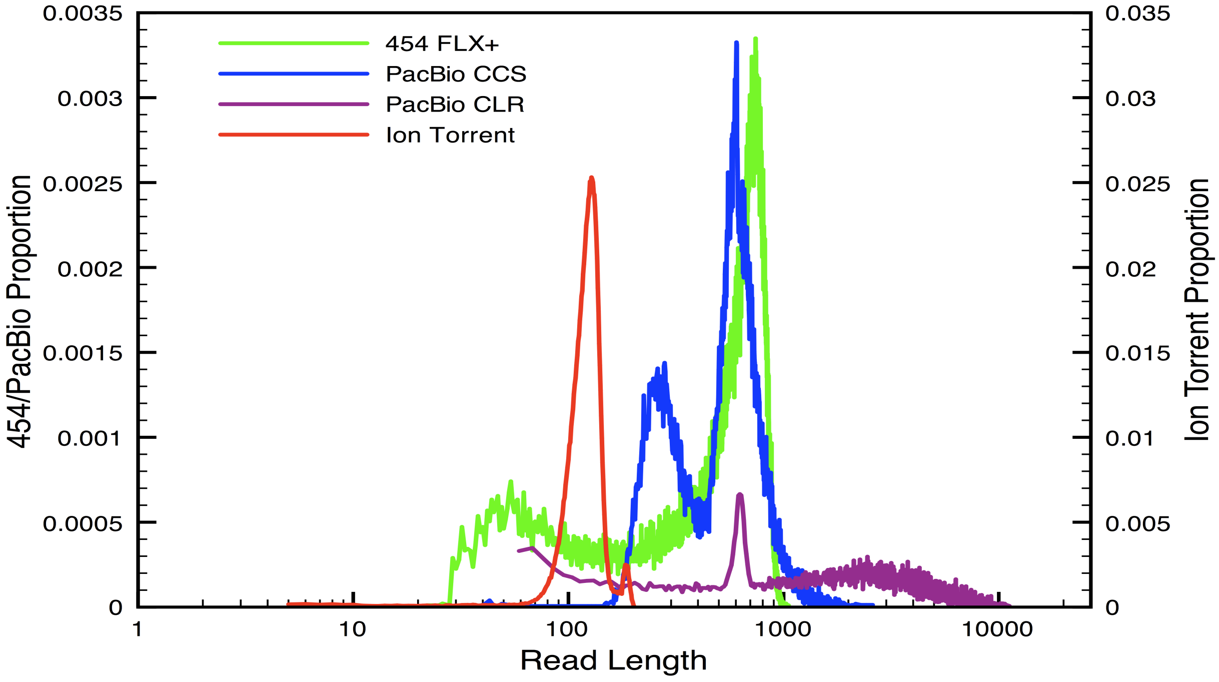
Is there a formula? Is there a work flow?

- Combining platforms, using read pairs, can produce a perfect genome in principle
- *Enterococcus faecalis* TX0309B model for R&D
- Data from
    - Illumina GAIIx pairs
    - Illumina MiSeq pairs
    - 454 FLX+ frags
    - Ion Torrent frags
    - PacBio
        - short (CCS lower error)
        - long (CLS 3kb avg; 13kb max)
        - Long reads lower accuracy; corrected with Illumina data
- Whole genome map using OpGen Argus technology

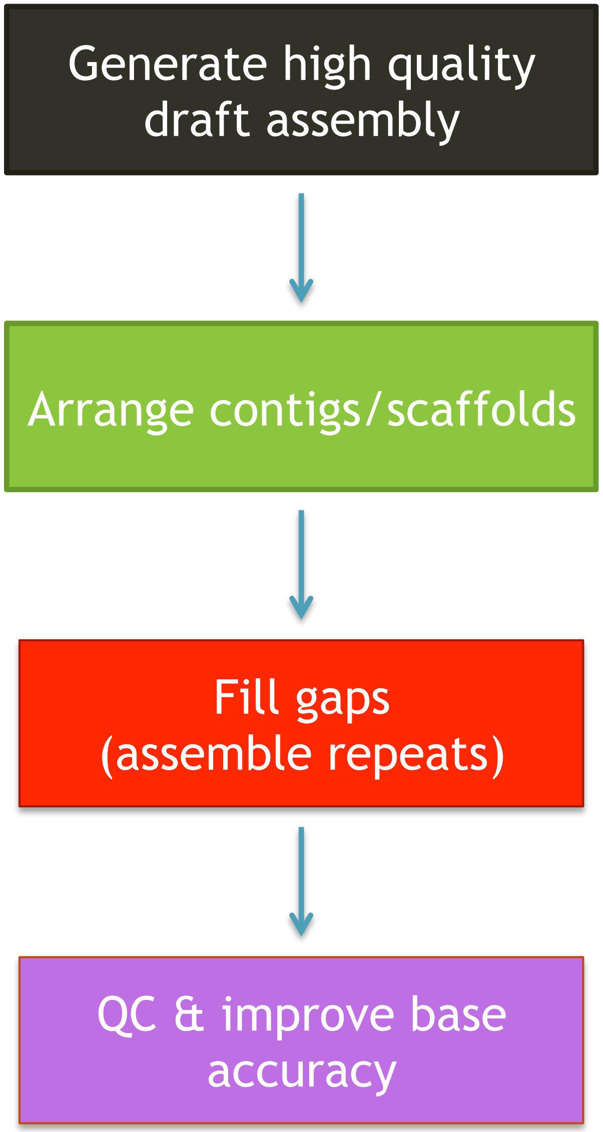
Vince Magrini, Jason Walker, Todd Wylie, Elaine Mardis

Enterococcus faecalis TX0309B Data

Technology	Platform	Library Type	Coverage - Type
PacBio	RS	CLR	97X - 10 Kbp Continuous Long Reads
PacBio	RS	CCS	31X - Circular Consensus Sequencing
Illumina	GAIIx	Paired-end	109X - Original HMP Velvet Assembly Data
Illumina	MiSeq	Mate-pair	254X – 3kb inserts
Illumina	MiSeq	Paired-end	464X - 170 bp inserts for overlapping "Sloptigs”
454	FLX+	Fragment	19X - 1500 bp library
IonTorrent	PGM	Fragment	29X - 100 bp library (314 and 316 Ion Chip)



Overall Strategy

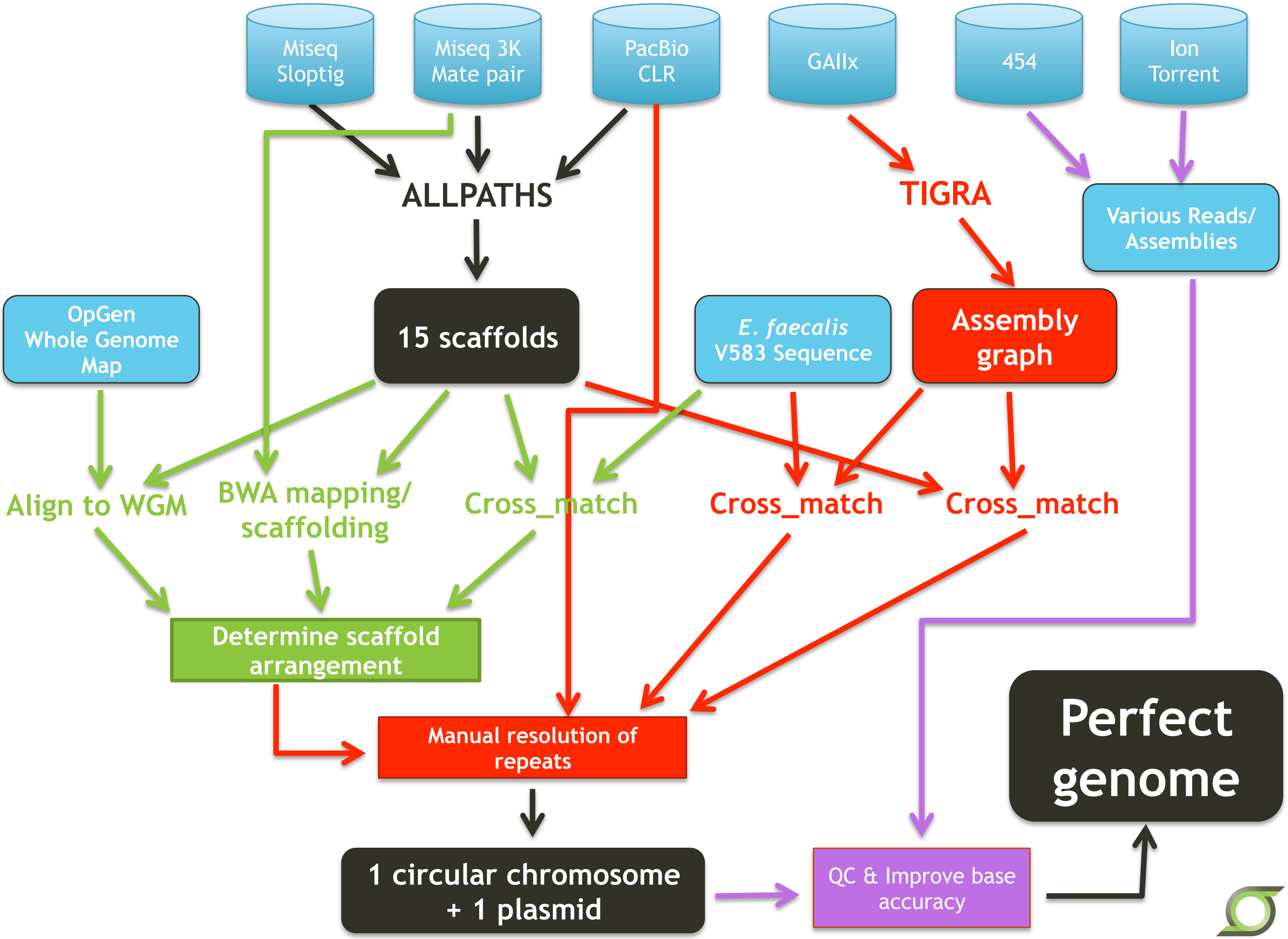


Generate high quality draft assembly

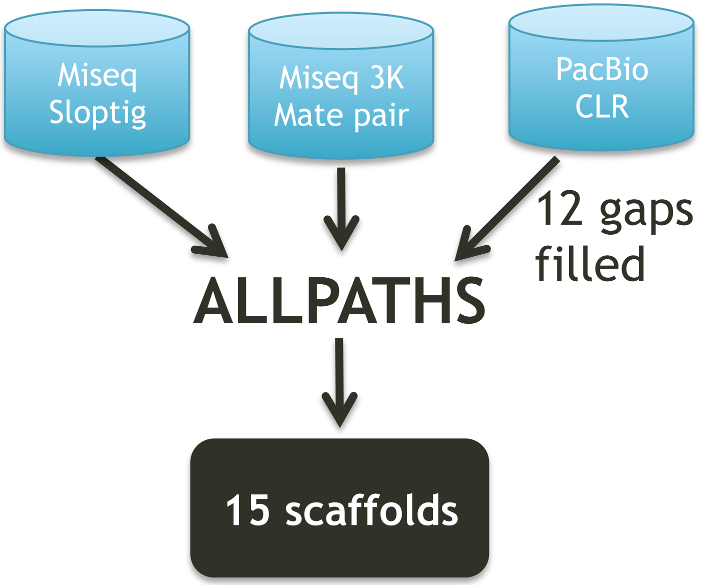
Arrange contigs/scaffolds

Fill gaps (assemble repeats)

QC & improve base accuracy


Miseq Sloptig
Miseq 3K Mate pair
PacBio CLR
GAIIx
454
Ion Torrent
ALLPATHS
TIGRA
Various Reads/ Assemblies
OpGen Whole Genome Map
15 scaffolds
E. faecalis V583 Sequence
Assembly graph
Align to WGM
BWA mapping/ scaffolding
Cross_match
Cross_match
Cross_match
Determine scaffold arrangement
Manual resolution of repeats
Perfect genome
1 circular chromosome + 1 plasmid
QC & Improve base accuracy

Generate high quality draft assembly

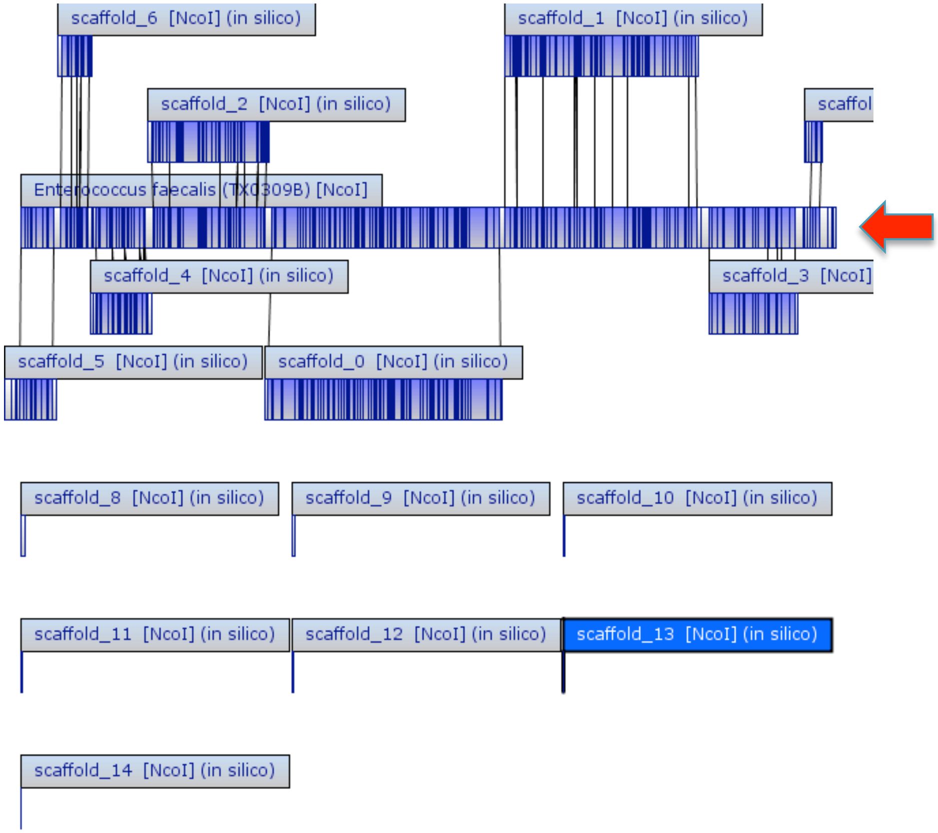


Scaffolds/contigs: 15
N50: 738,922
Num_to_N50: 2
Total length: 3,137,099
Mean: 209,139
Max: 907,745

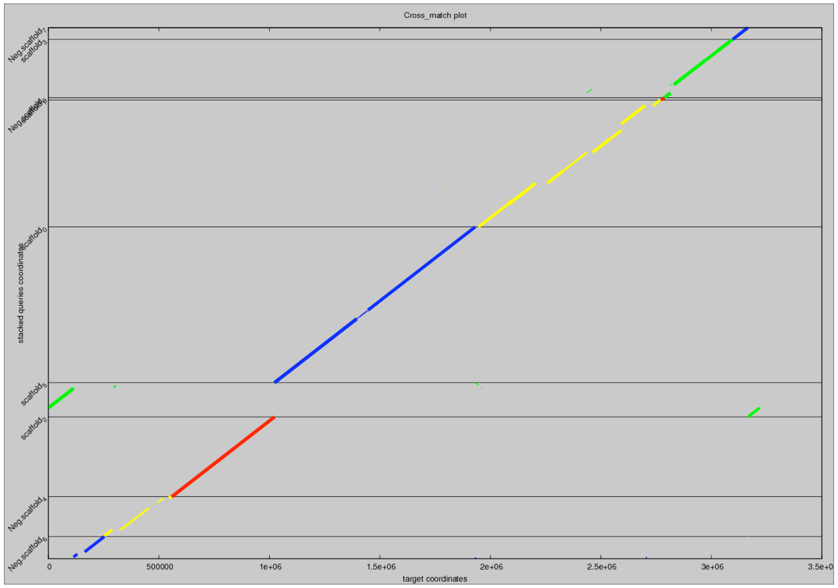
E. faecalis TX0309B Draft Genome

- Various reads, various assemblers (Newbler, Velvet, Celera, ALLPATHS) tested
- ALLPATHS best assembly so far; 15 scaffolds (each single contig)
- Whole genome map used for QA at this stage

Align to Whole Genome Map


scaffold_6 [NcoI] (in silico)
scaffold_1 [NcoI] (in silico)
scaffold_2 [NcoI] (in silico)
scaffol
Enterococcus faecalis (TX0309B) [NcoI]
scaffold_4 [NcoI] (in silico)
scaffold_3 [NcoI]
scaffold_5 [NcoI] (in silico)
scaffold_0 [NcoI] (in silico)
scaffold_8 [NcoI] (in silico)
scaffold_9 [NcoI] (in silico)
scaffold_10 [NcoI] (in silico)
scaffold_11 [NcoI] (in silico)
scaffold_12 [NcoI] (in silico)
scaffold_13 [NcoI] (in silico)
scaffold_14 [NcoI] (in silico)

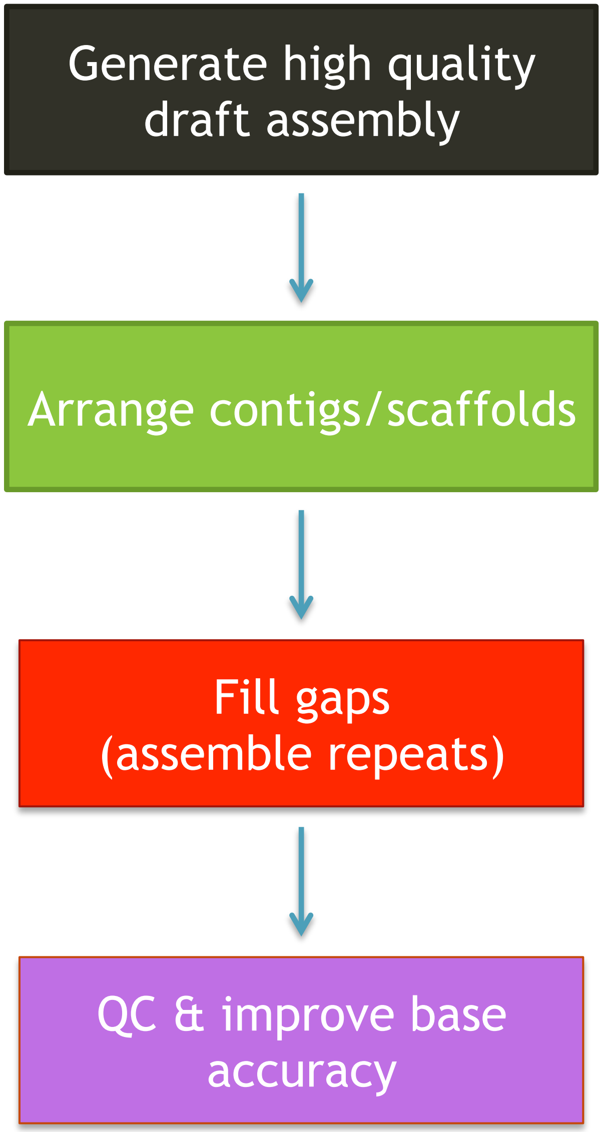
Align to *E. faecalis* V583



The TX0309B Chromosome & Repeats

- 1 chromosome and 1 plasmid
- Chromosome: 10 scaffolds, due to 9 repeats
  - 8 of which mapped in the Whole Genome Map (OpGen)
  - 9 of which mapped to V583
- 9 repeats are from 4 repeat types
  - 18 kb transposon, appeared twice
  - Tandem repeat of a 300 bp unit, 9 to 15+ tandem copies, pathogenicity island
  - 5 kb rRNA complex, repeated 4 times (complex ALLPATHS structure)
  - 15 kb phage insertion, repeated twice
- Plasmid: 5 ALLPATHS scaffolds

Overall Strategy

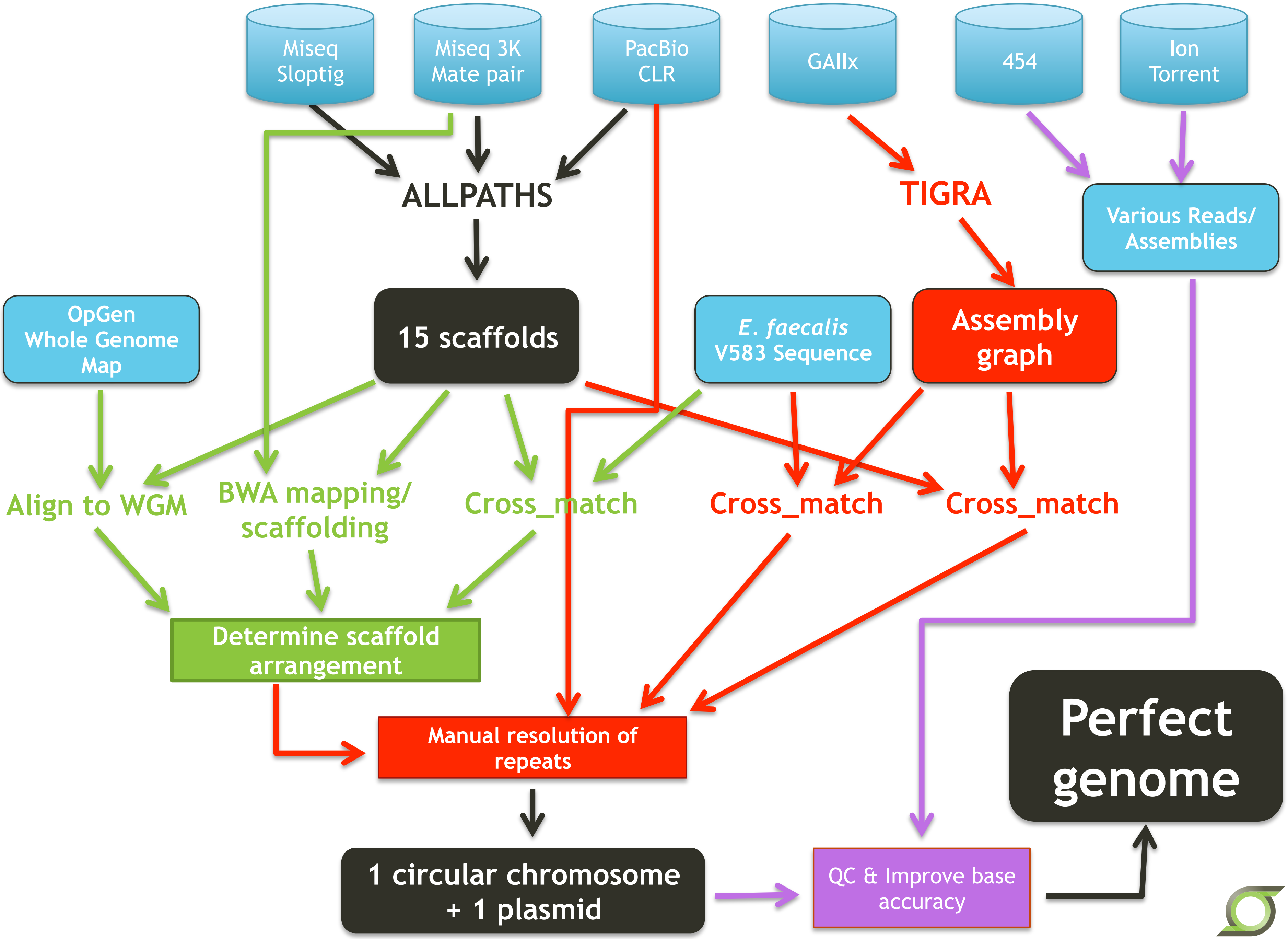


Generate high quality draft assembly

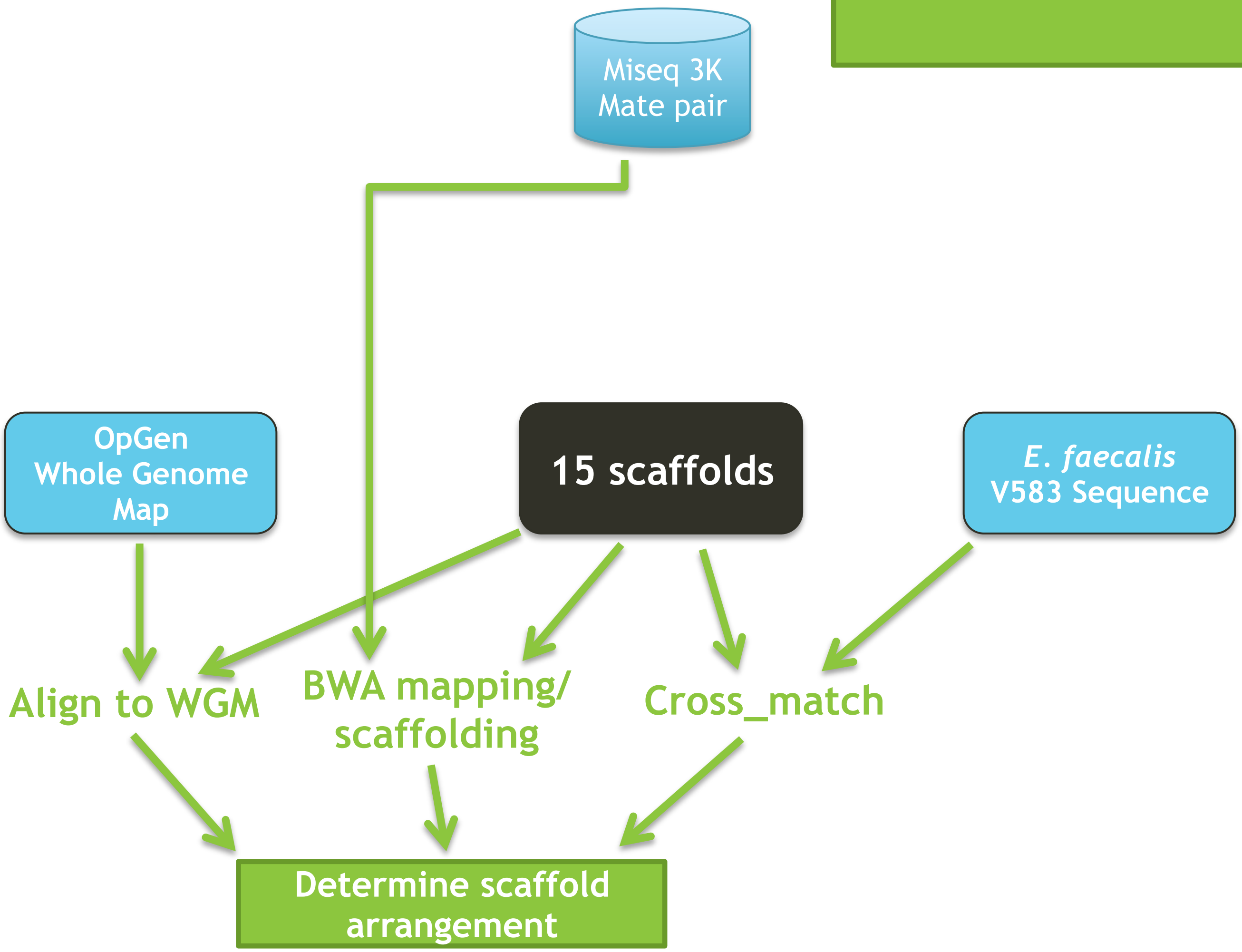
Arrange contigs/scaffolds

Fill gaps (assemble repeats)

QC & improve base accuracy


Miseq Sloptig
Miseq 3K Mate pair
PacBio CLR
GAIIx
454
Ion Torrent
ALLPATHS
TIGRA
Various Reads/ Assemblies
OpGen Whole Genome Map
15 scaffolds
E. faecalis V583 Sequence
Assembly graph
Align to WGM
BWA mapping/ scaffolding
Cross_match
Cross_match
Cross_match
Determine scaffold arrangement
Manual resolution of repeats
Perfect genome
1 circular chromosome + 1 plasmid
QC & Improve base accuracy

Arrange contigs/scaffolds



Miseq 3K Mate pair

OpGen Whole Genome Map

15 scaffolds

E. faecalis V583 Sequence

Align to WGM

BWA mapping/ scaffolding

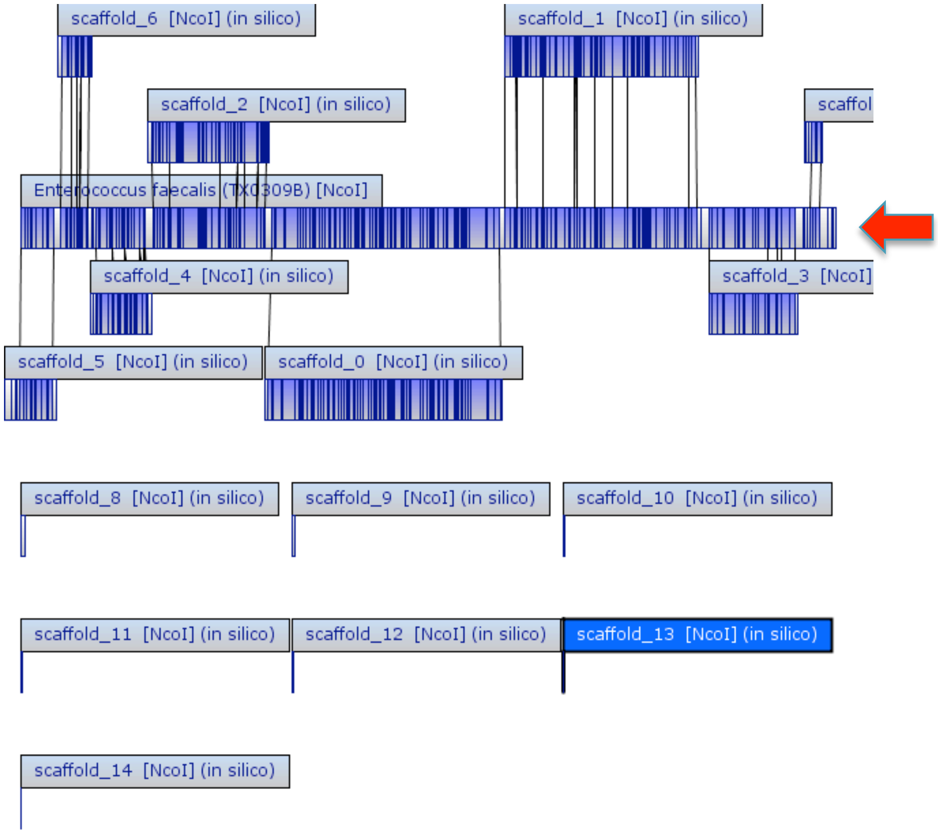
Cross_match

Determine scaffold arrangement

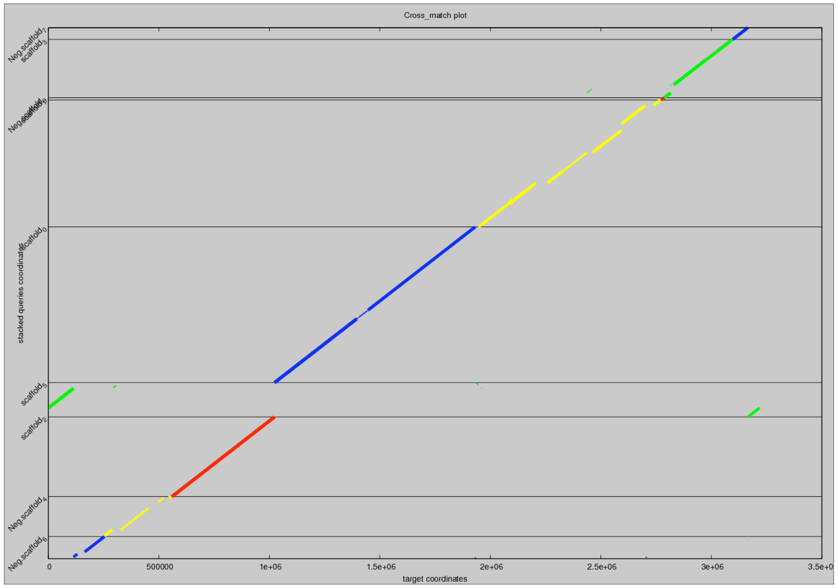
E. faecalis TX0309B Draft Genome

- Various reads, various assemblers (Newbler, Velvet, Celera, ALLPATHS) tested
- ALLPATHS best assembly so far; 15 scaffolds (each single contig)
- Whole genome map used for QA at this stage

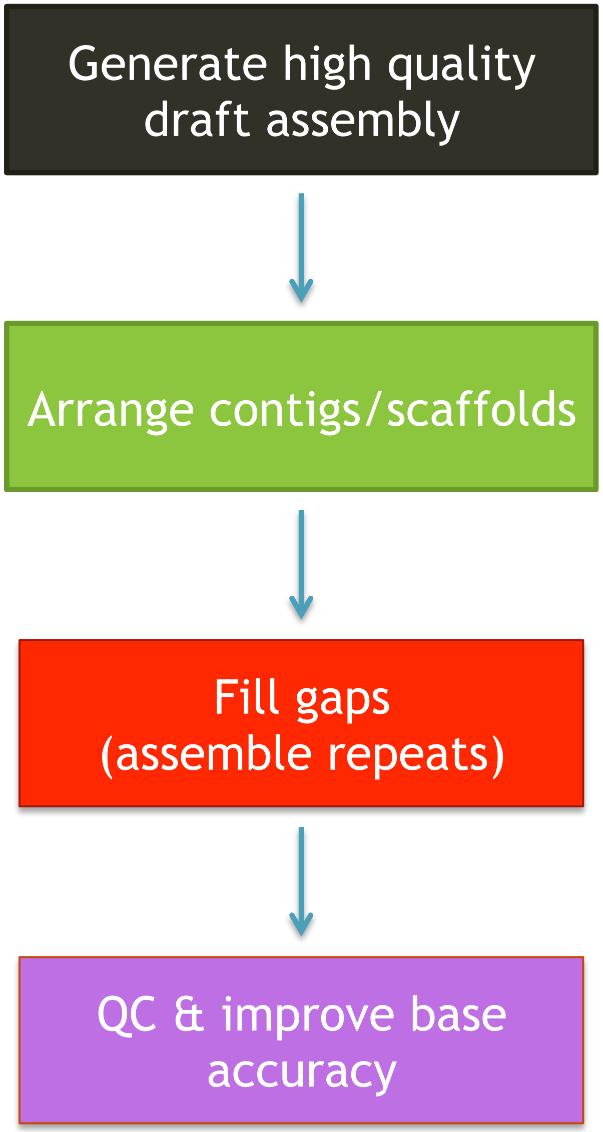
Align to Whole Genome Map


scaffold_6 [NcoI] (in silico)
scaffold_1 [NcoI] (in silico)
scaffold_2 [NcoI] (in silico)
scaffol
Enterococcus faecalis (TX0309B) [NcoI]
scaffold_4 [NcoI] (in silico)
scaffold_3 [NcoI]
scaffold_5 [NcoI] (in silico)
scaffold_0 [NcoI] (in silico)
scaffold_8 [NcoI] (in silico)
scaffold_9 [NcoI] (in silico)
scaffold_10 [NcoI] (in silico)
scaffold_11 [NcoI] (in silico)
scaffold_12 [NcoI] (in silico)
scaffold_13 [NcoI] (in silico)
scaffold_14 [NcoI] (in silico)

Align to *E. faecalis* V583



Overall Strategy

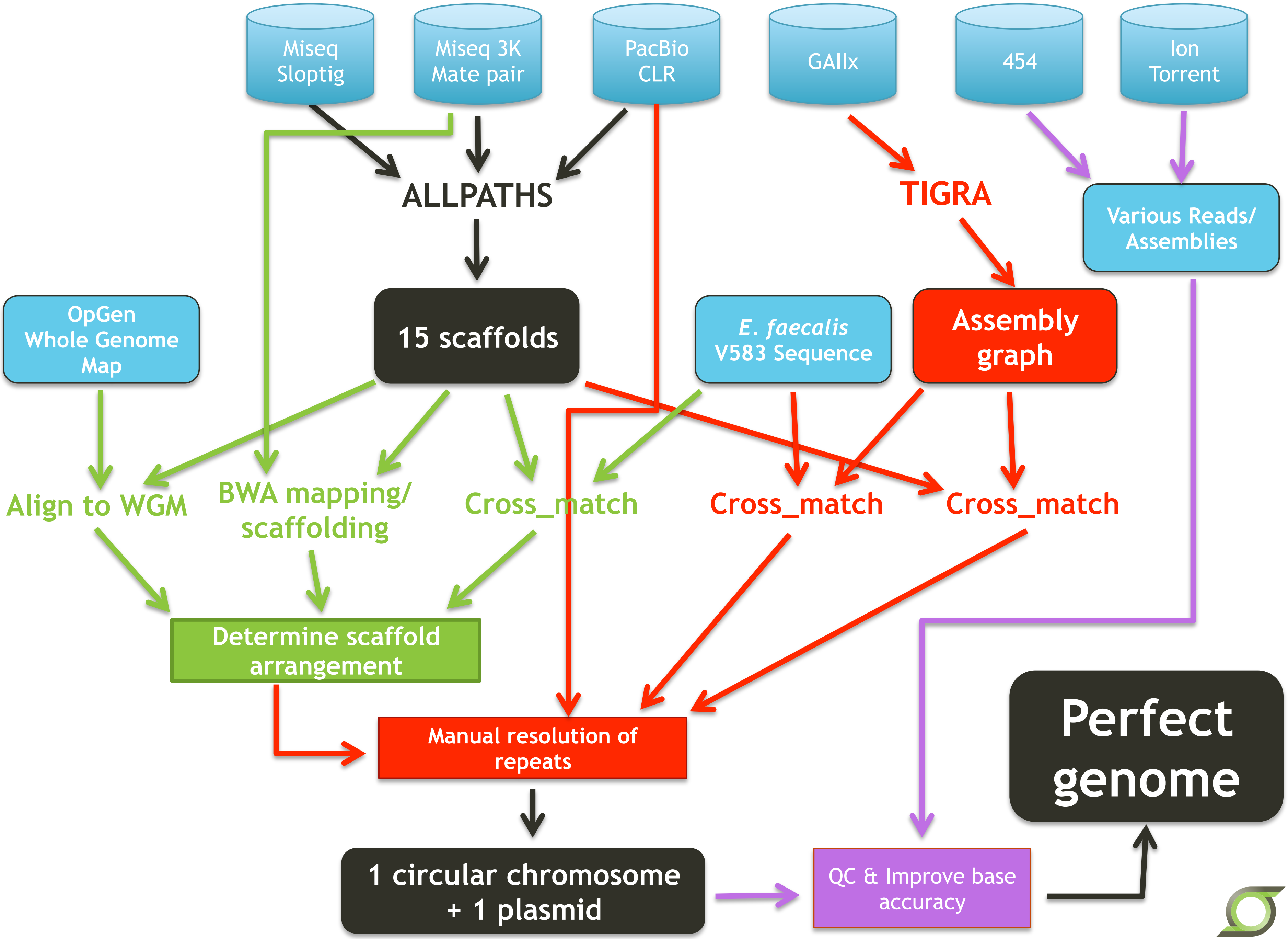


Generate high quality draft assembly

Arrange contigs/scaffolds

Fill gaps (assemble repeats)

QC & improve base accuracy



Miseq Sloptig

Miseq 3K Mate pair

PacBio CLR

GAIIx

454

Ion Torrent

ALLPATHS

TIGRA

Various Reads/ Assemblies

OpGen Whole Genome Map

15 scaffolds

E. faecalis V583 Sequence

Assembly graph

Align to WGM

BWA mapping/ scaffolding

Cross_match

Cross_match

Cross_match

Determine scaffold arrangement

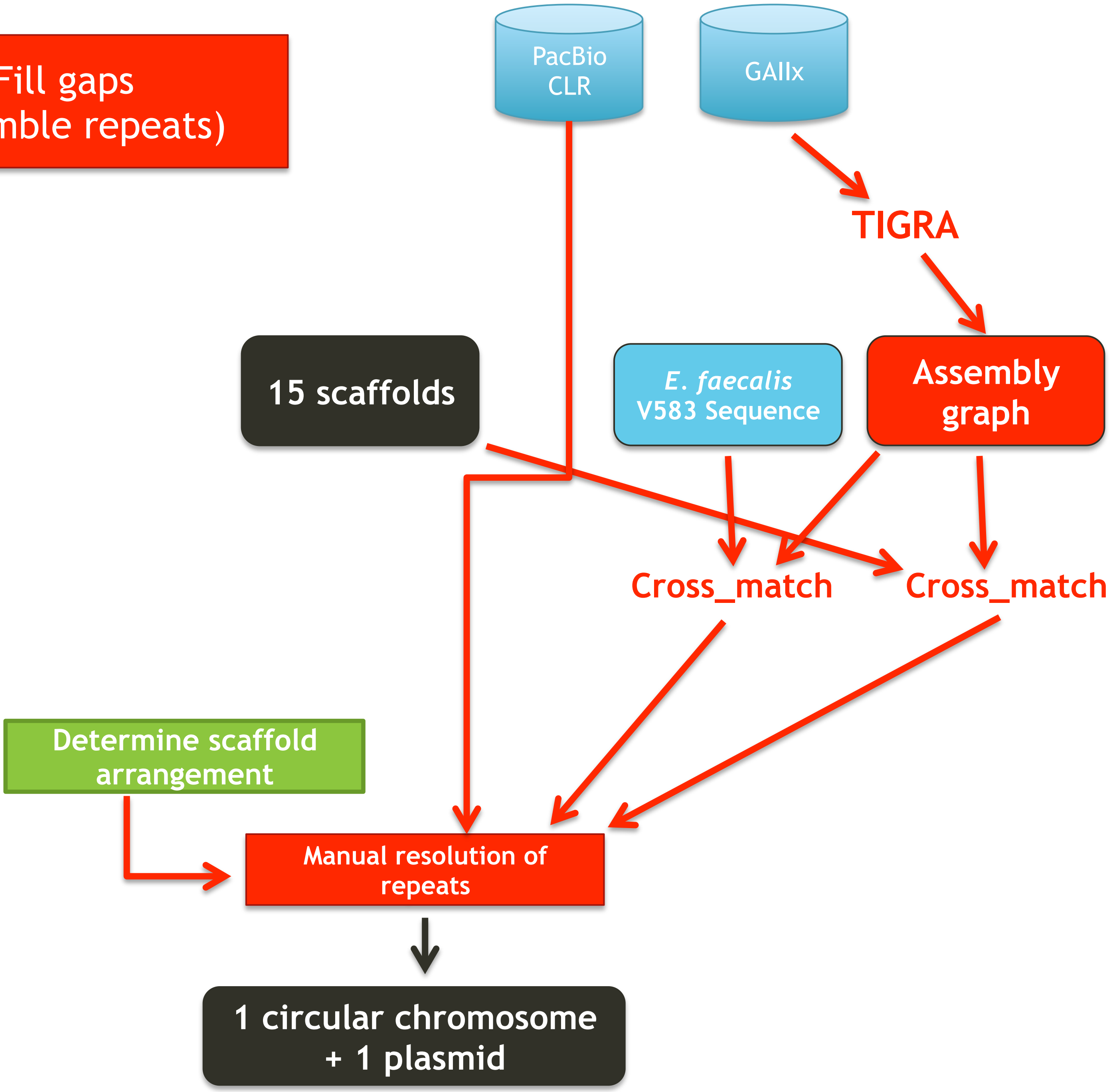
Manual resolution of repeats

Perfect genome

1 circular chromosome + 1 plasmid

QC & Improve base accuracy

Fill gaps (assemble repeats)



PacBio CLR

GAIIx

TIGRA

15 scaffolds

E. faecalis V583 Sequence

Assembly graph

Cross_match

Cross_match

Determine scaffold arrangement

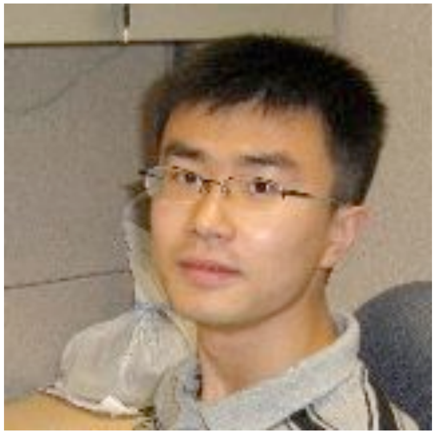
Manual resolution of repeats

1 circular chromosome + 1 plasmid

Why TIGRA?

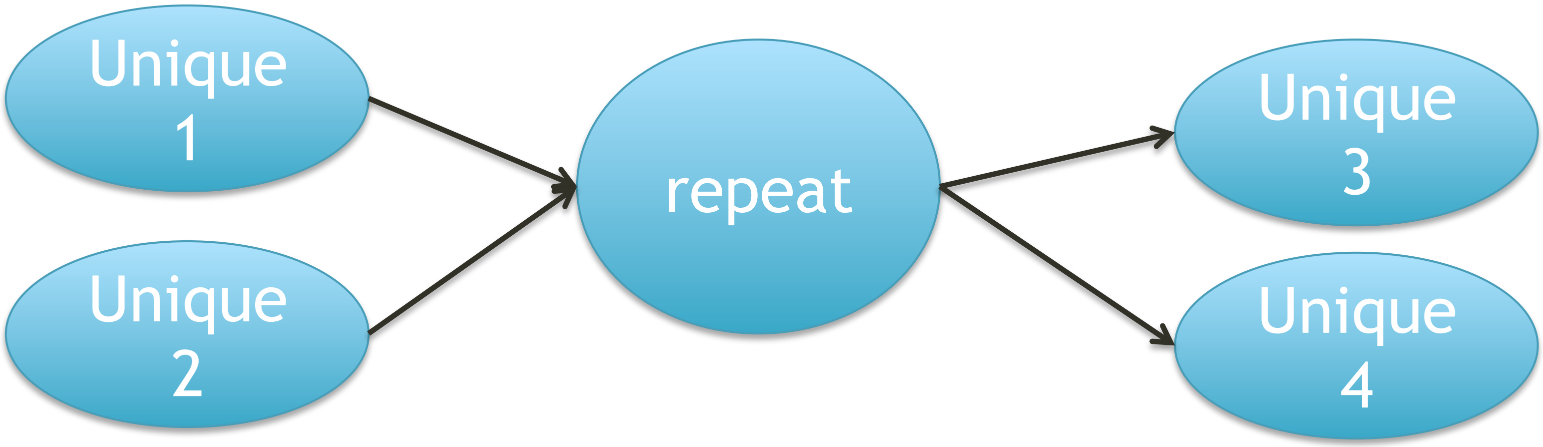
- TIGRA: The Iterative Graph Routine Assembler
- Based on DeBruijn graph assemblers
- Does not require particular types of libraries, platforms
- TIGRA preserves unresolved repeats and polymorphisms
  - Does NOT make links to gain bigger N50
  - Does NOT collapse repeats or variants into single consensus
- Supports the assembly graph visualization
- Extensively used for analyzing structural variation in human genomes

Lei Chen

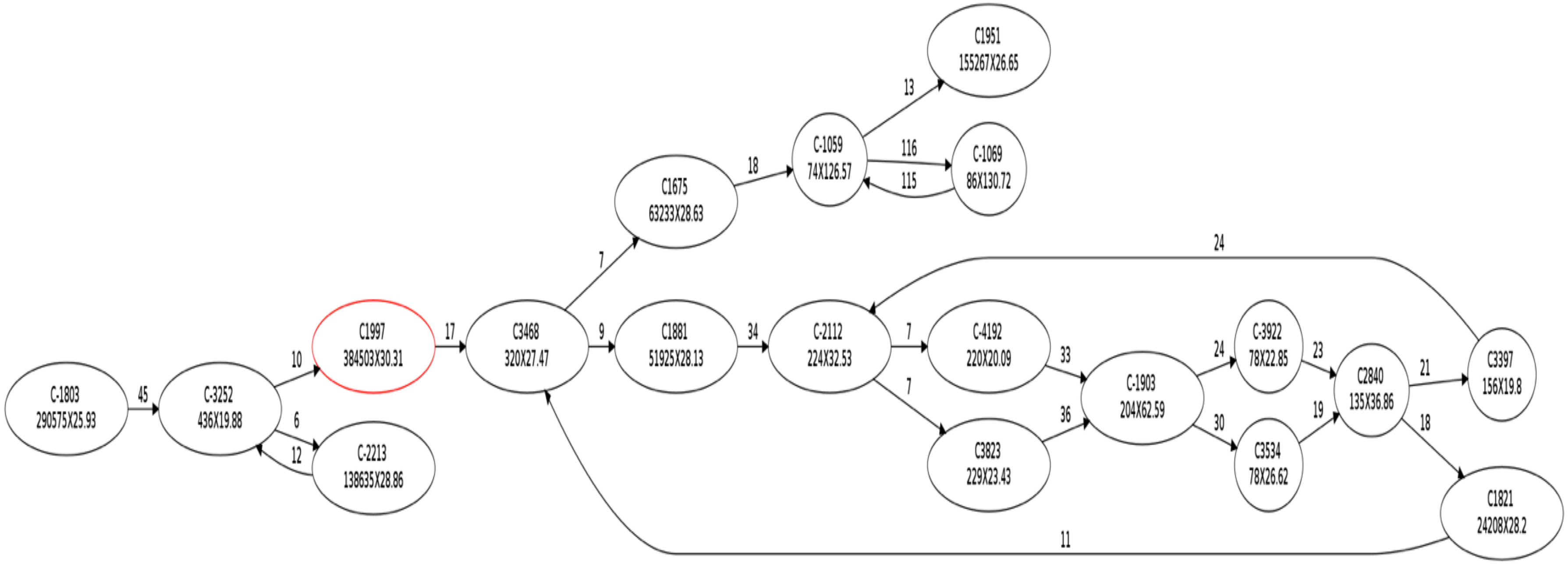


Assembly Graph

- Shows the links and order between contigs/scaffolds
- Nodes: contigs/scaffolds
- Edges: directed



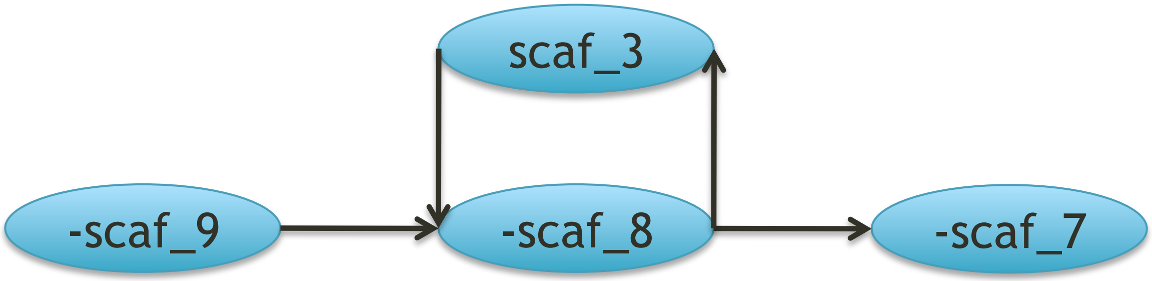
Genome of *Rickettsia prowazekii*



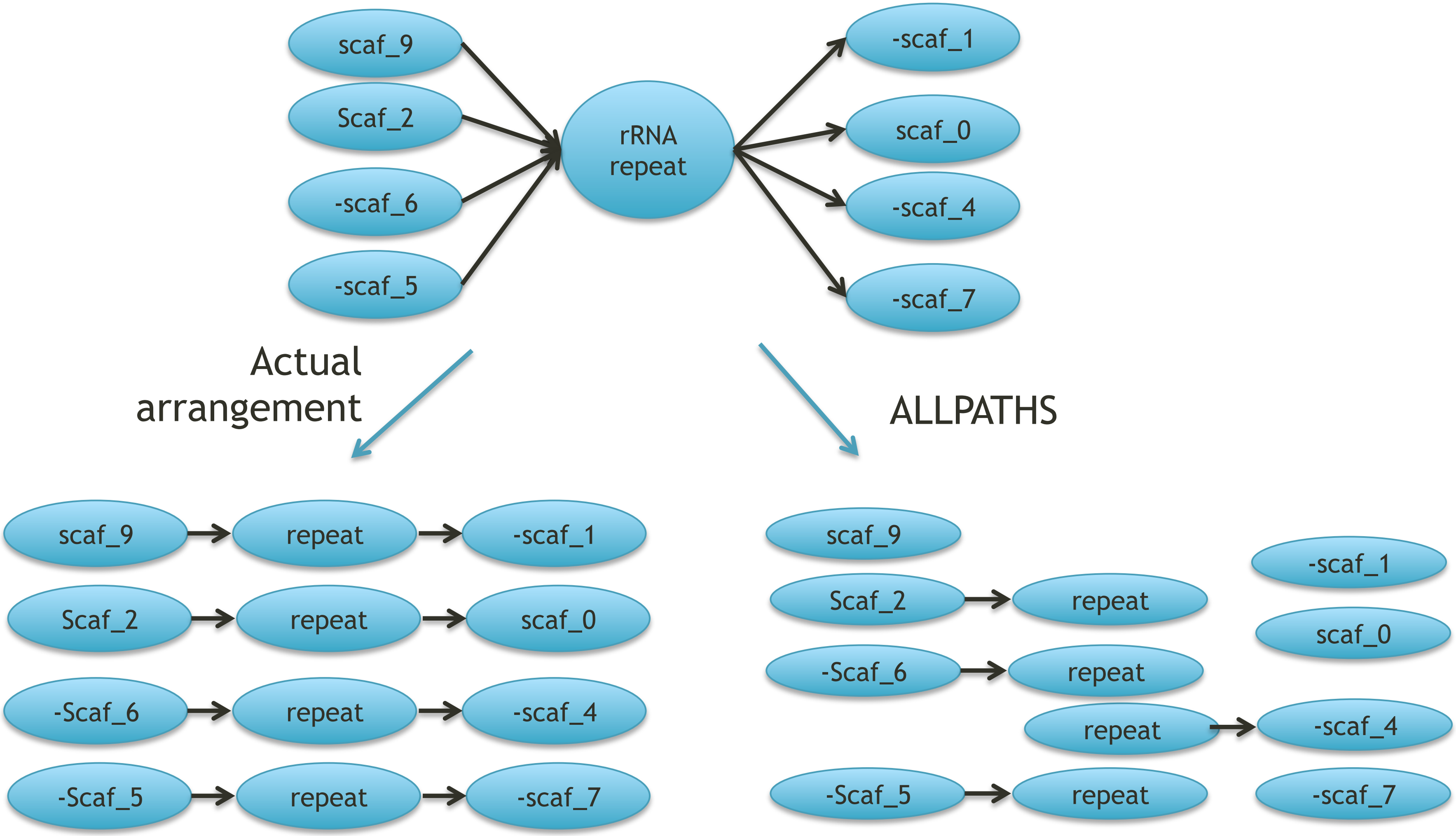
Assembled with TIGRA

The 18kb transposon

- Scaf8 is the transposon sequence that’s not present in V583.
- ALLPATHS did allright with this repeat : one scaffold by itself.
- Used mate pair mapping, coverage analysis, and rough gap estimate from WGM



Mis-assembly in ALLPATHS, the rRNA example



scaf_9 → rRNA repeat

Scaf_2 → rRNA repeat

-scaf_6 → rRNA repeat

-scaf_5 → rRNA repeat

rRNA repeat → -scaf_1

rRNA repeat → scaf_0

rRNA repeat → -scaf_4

rRNA repeat → -scaf_7

Actual arrangement

scaf_9 → repeat → -scaf_1

Scaf_2 → repeat → scaf_0

-Scaf_6 → repeat → -scaf_4

-Scaf_5 → repeat → -scaf_7

ALLPATHS

scaf_9

-scaf_1

Scaf_2 → repeat

scaf_0

-Scaf_6 → repeat

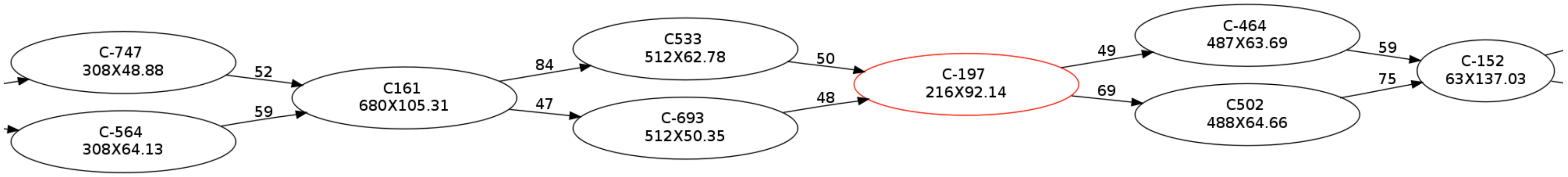
repeat → -scaf_4

-Scaf_5 → repeat

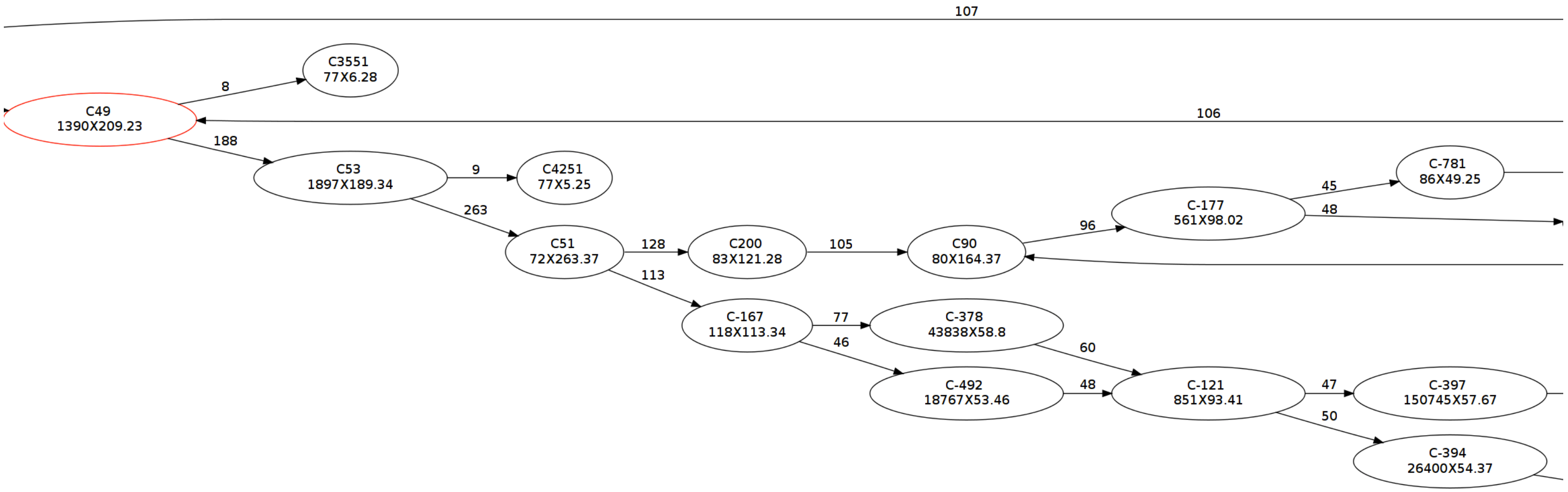
-scaf_7

Assembly Graph by TIGRA

Part of the
phage repeat



Part of the
rRNA repeat



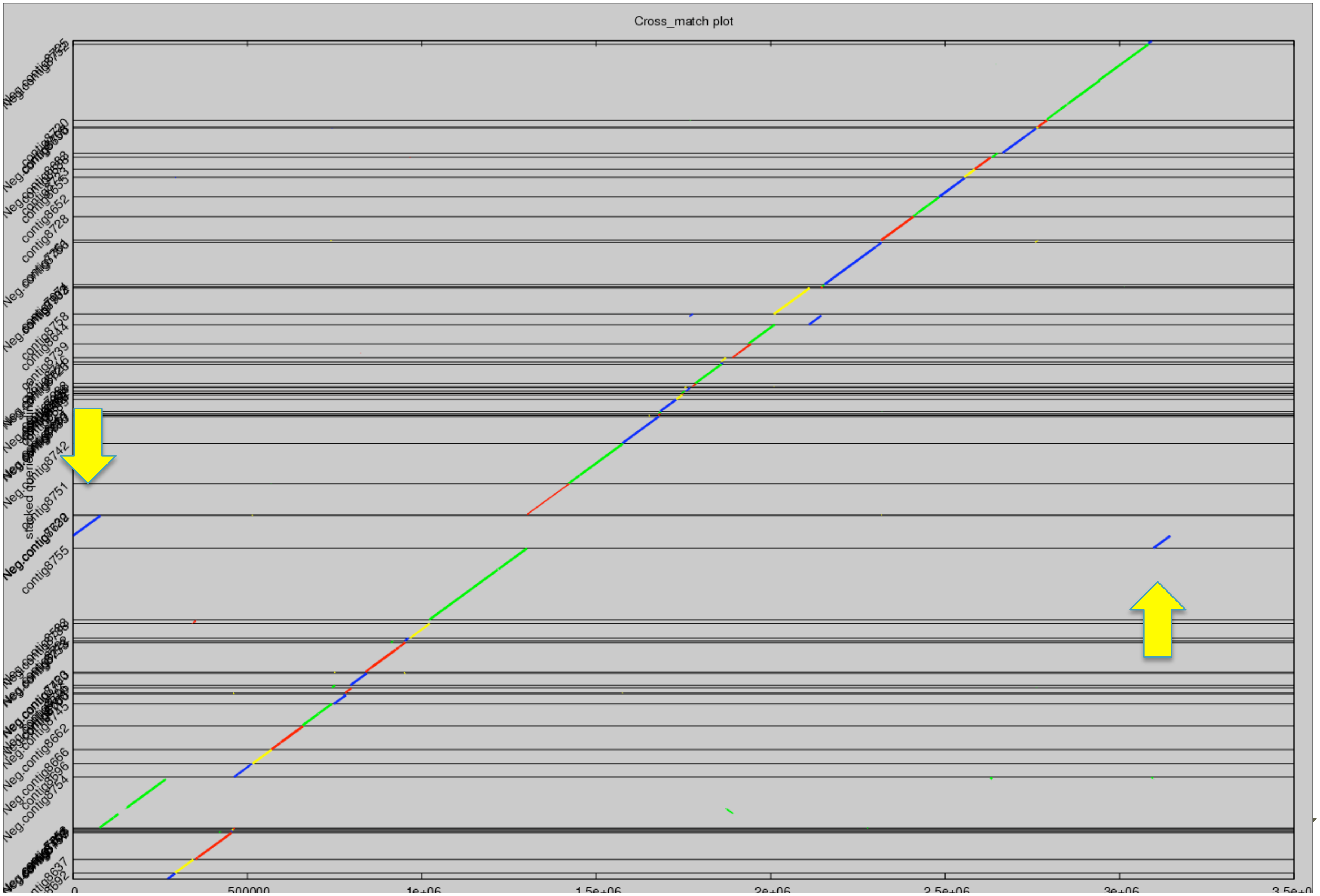
Repeat Assembly

- Repeat are not perfect, fragment size limit how big a repeat can be resolved
- V583 alignment used to determine where different versions of repeats should be placed
  - Rely on small sequence differences
- PacBio CLR used to determine copy number of the tandem repeat.

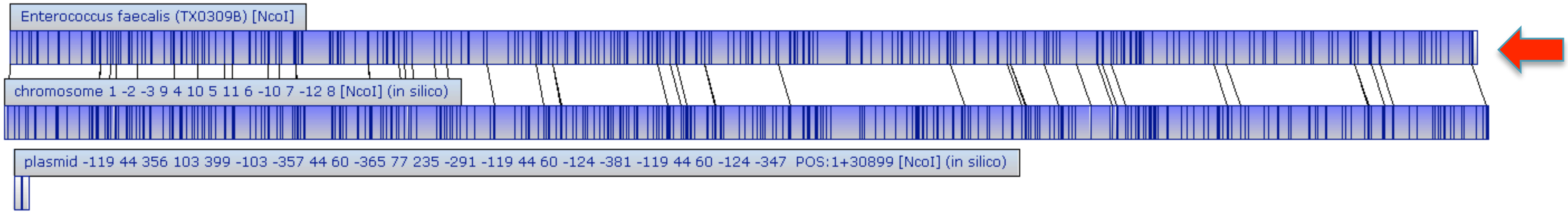
Assembly QC

- Align to existing assemblies to check mis-assembly
  - 454 by newbler
    PacBio by allora
    GAIIx by velvet
    Ion Torrent by newbler
- All assemblies gave contiguous alignment of each contig except PacBio
  - Some allora contigs were split indicating they were mis-assembled

PacBio has split contigs



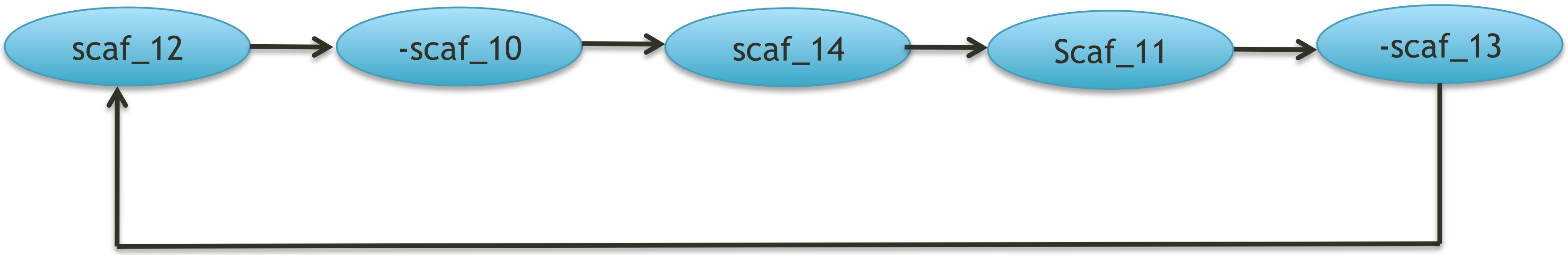
The TX0309B genome aligns to the WGM


Enterococcus faecalis (TX0309B) [NcoI]
chromosome 1 -2 -3 9 4 10 5 11 6 -10 7 -12 8 [NcoI] (in silico)
plasmid -119 44 356 103 399 -103 -357 44 60 -365 77 235 -291 -119 44 60 -124 -381 -119 44 60 -124 -347 POS:1+30899 [NcoI] (in silico)

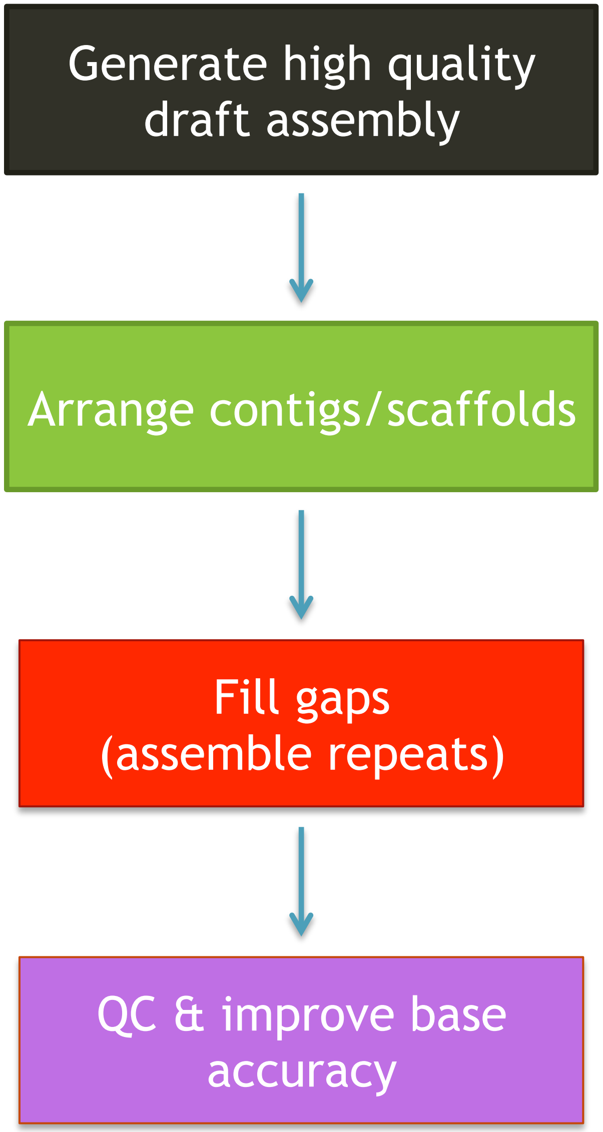
Gaps are correctly filled.
The contigs/scaffolds are correctly aligned.
The assembly is in a single contig (+ a plasmid).

The plasmid

- One circular plasmid
- Consists of 5 ALLPATHS scaffolds
- Closed by alignment to TIGRA assembly
- Confirmed by mate pair mapping



Overall Strategy

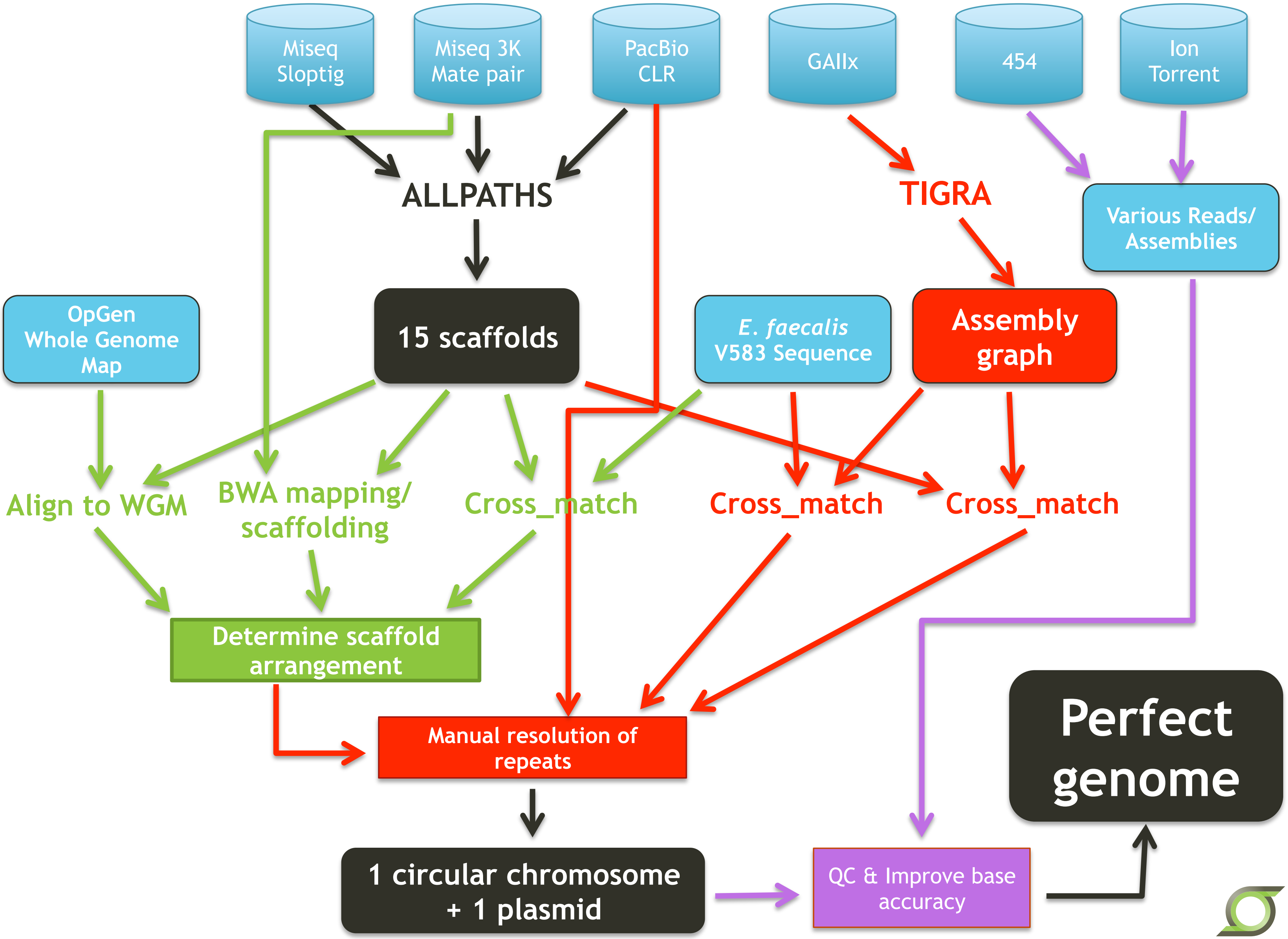


Generate high quality draft assembly

Arrange contigs/scaffolds

Fill gaps (assemble repeats)

QC & improve base accuracy



Miseq Sloptig

Miseq 3K Mate pair

PacBio CLR

GAIIx

454

Ion Torrent

ALLPATHS

TIGRA

Various Reads/ Assemblies

OpGen Whole Genome Map

15 scaffolds

E. faecalis V583 Sequence

Assembly graph

Align to WGM

BWA mapping/ scaffolding

Cross_match

Cross_match

Cross_match

Determine scaffold arrangement

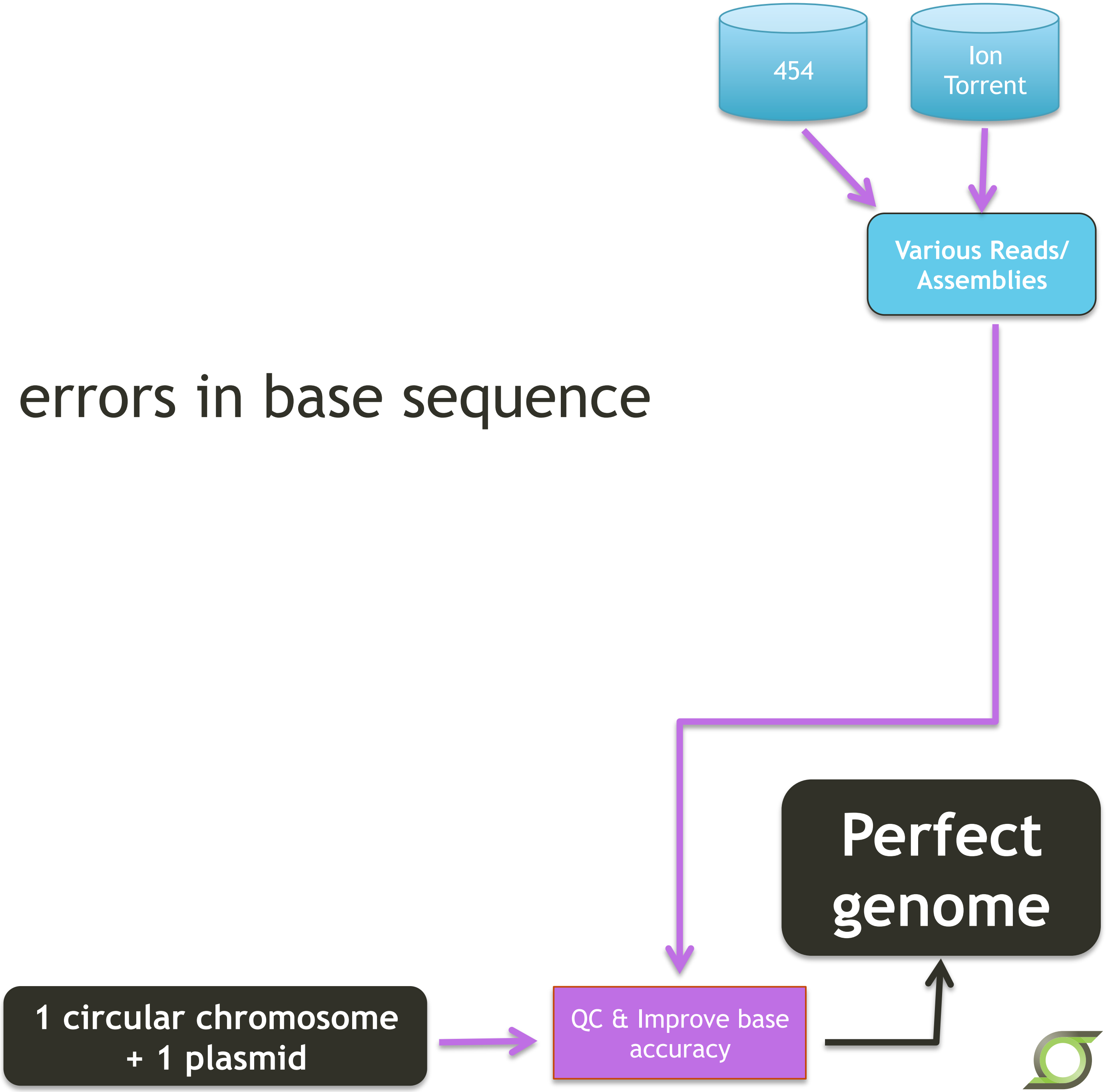
Manual resolution of repeats

Perfect genome

1 circular chromosome + 1 plasmid

QC & Improve base accuracy

Correcting errors in base sequence


454
Ion Torrent
Various Reads/ Assemblies
Perfect genome
1 circular chromosome + 1 plasmid
QC & Improve base accuracy

Map reads and call variants to improve base accuracy

- Due to the large variety of data available, it’s still a work in progress.
- ALLPATHS assembly contains ambiguous base code: R, Y … 357 in total. These are treated as N by many variant callers.
- Some of these ambiguous bases are valid polymorphisms.
- Some are due to repeats treated as polymorphic regions.

Current Mapping & Variant Calling Status

Platform	Mapping Software	# of variants called	Note
454 FLX+ Frag	Newbler runMapping	383	
GAIIx	BWA/Samtools	405	
MiSeq 3kb mate pair (not used by ALLPATHS)	BWA/Samtools	450	
Ion torrent paired	BWA/Samtools	2903	Many homopolymeric indels
PacBio CCS	BWA-SW/Samtools	13540	Many homopolymeric indels

Consensus Calling

- Whenever there are at least 3 sources (ie 3/5) indicating the same variant, it’s changed accordingly.
- Made 361 changes in total, 301 of them are for ALLPATHS ambiguous bases.

Conclusion

- Expectation is that a “perfect” genome can be achieved
  - Much higher quality than current “finished”
  - Will need a combination of 2 (or more?) NGS platforms and whole genome map
- Faster, cheaper, higher quality than current Gold Standard genomes
- No (Sanger) finishing required (?)

Acknowledgments

- NGS data production by Vince Magrini and Elaine Mardis, Technology Development group staff
- Data processing and management by Jason Walker and Todd Wylie, Technology Development group staff
- Whole Genome Map: Amy Ly, Technology Development group
- Lei Chen: TIGRA and analysis
- Guohui Yao: PyGap and Pyramid
- Microbial Genomics group: Erica Sodergren et al.
- *Treponema pallidum* over the years: David Šmajs et al. (Masaryk Univ., Brno)