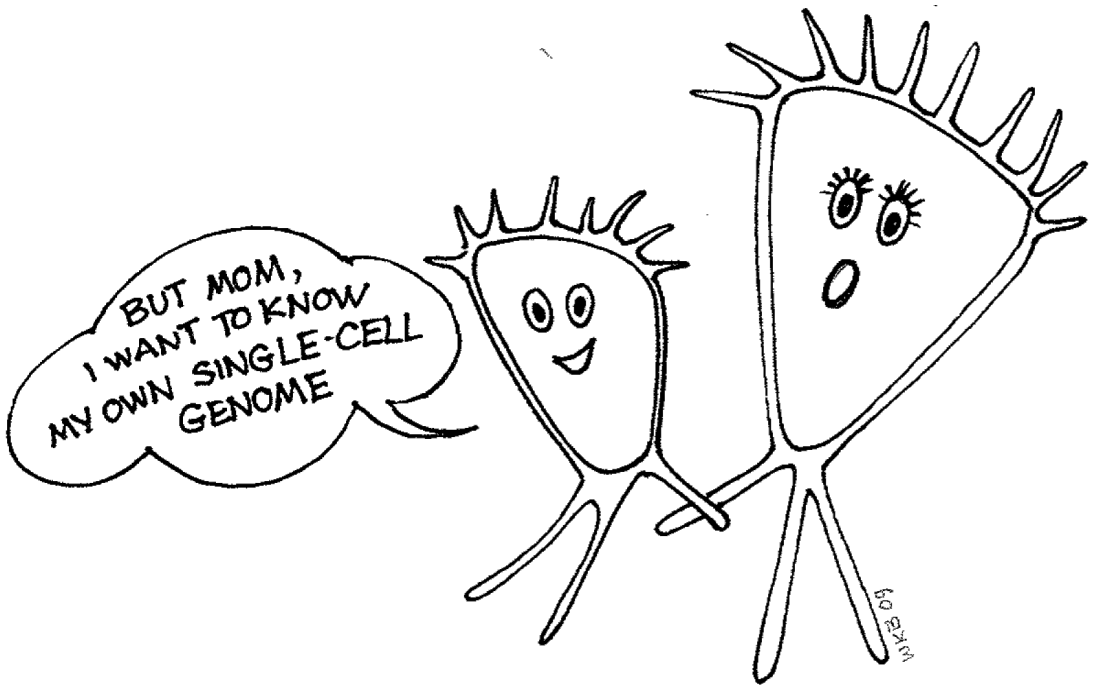
Bigelow | Laboratory for Ocean Sciences

Ramunas Stepanauskas



SAG quality control

Current QC reporting standards

nature biotechnology
OPEN

PERSPECTIVE

Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea

Table 1 Genome reporting st

Criterion	
Finished	
Assembly quality[a]	or better
High-quality draft (SAG/MAG)	
Assembly quality[a]	Multiple fragments where gaps span repetitive regions. Presence of the 23S, 16S, and 5S rRNA genes and at least 18 tRNAs.
Completion[b]	>90%
Contamination[c]	<5%
Medium-quality draft	
Assembly	
Comple	
Contamination	
Low-quality draft (SAG/MAG)	
Assembly quality[a]	Many fragments with little to no review of assembly other than reporting of standard assembly statistics.
Completion[b]	<50%
Contamination[c]	<10%

This is a compre

[a]Assem dicted genes per g opy marker genes in ≥2 copies to tota

Problem #1: Contamination limits too permissive

Problem #2: Misassemblies, indels, mismatches not covered

Problem #3: QC method validation not covered

Bigelow | Laboratory for Ocean Sciences

Current QC reporting standards

Method

CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes

Donovan H. Parks,[1] Michael Imelfort,[1] Connor T. Skennerton,[1] Philip Hugenholtz,[1,2] and Gene W. Tyson[1,3]

[1]Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, St. Lucia, QLD 4072, Queensland, Australia; [2]Institute for Molecular Bioscience, The University of Queensland, St. Lucia, QLD 4072, Queensland, Australia; [3]Advanced Water Management Centre, The University of Queensland, St. Lucia, QLD 4072, Queensland, Australia

Contamination under-estimation by checkM

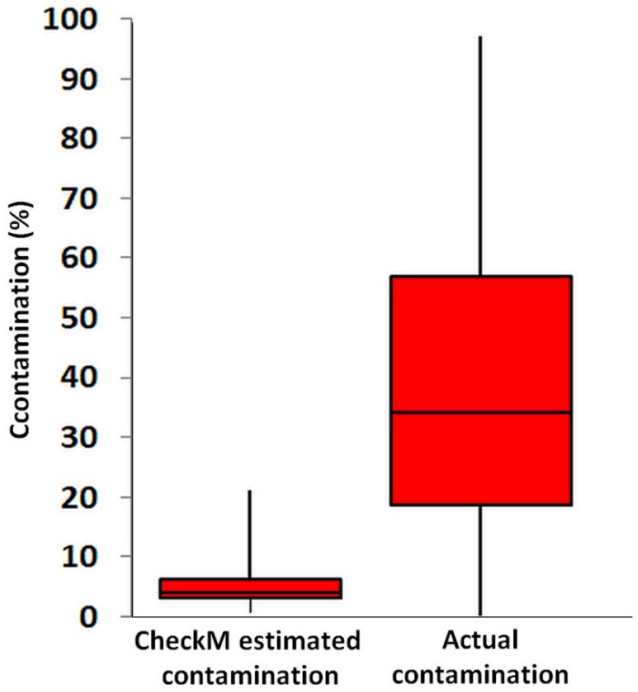


FIGURE 6 | Contamination predicted by CheckM software from all pairwise SAG combinations across the class-level lineages Rokumicrobia and Infratellusbacteria (15 Rokumicrobia and 4 Infratellusbacteria SAGs; 60 combinations total), compared to actual contamination calculated from all artificially combined SAGs. Two-sample *t*-test assuming equal variances was significant ($p = <000.1$).

From: Becraft et al. 2017

Reasons behind this discrepancy:

1. CheckM searches for multiple occurrences of same marker gene, not phylogeny conflicts.
2. Assembly incompleteness is not accounted for.
3. Reference database limitations may lead to detection failures of some marker genes .

Recent microbial genome quality re-evaluation

Orakov *et al. Genome Biology* (2021) 22:178
https://doi.org/10.1186/s13059-021-02393-0

Genome Biology

METHOD Open Access

GUNC: detection of chimerism and contamination in prokaryotic genomes

Askarbek Orakov[1†], Anthony Fullam[1†], Luis Pedro Coelho[2,3], Supriya Khedkar[1], Damian Szklarczyk[4,5], Daniel R. Mende[6], Thomas S. B. Schmidt[1*] and Peer Bork[1,7,8,9*]

* Correspondence: sebastian.schmidt@embl.de; peer.bork@embl.org
†Askarbek Orakov and Anthony Fullam contributed equally to this work.
[1]Structural and Computational Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany
Full list of author information is available at the end of the article

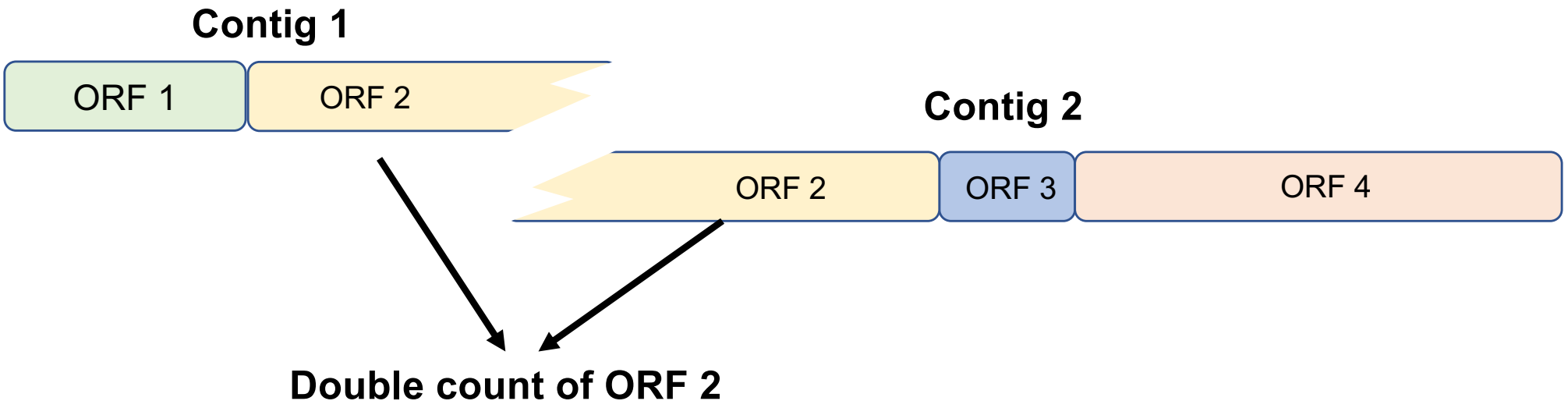
Abstract

Genomes are critical units in microbiology, yet ascertaining quality in prokaryotic genome assemblies remains a formidable challenge. We present GUNC (the Genome UNClutterer), a tool that accurately detects and quantifies genome chimerism based on the lineage homogeneity of individual contigs using a genome's full complement of genes. GUNC complements existing approaches by targeting previously underdetected types of contamination: we conservatively estimate that 5.7% of genomes in GenBank, 5.2% in RefSeq, and 15–30% of pre-filtered "high-quality" metagenome-assembled genomes in recent studies are undetected chimeras. GUNC provides a fast and robust tool to substantially improve prokaryotic genome quality.

Keywords: Genome quality, Genome contamination, Metagenomics, Metagenome-assembled genomes, Bioinformatics

Bigelow | Laboratory for Ocean Sciences

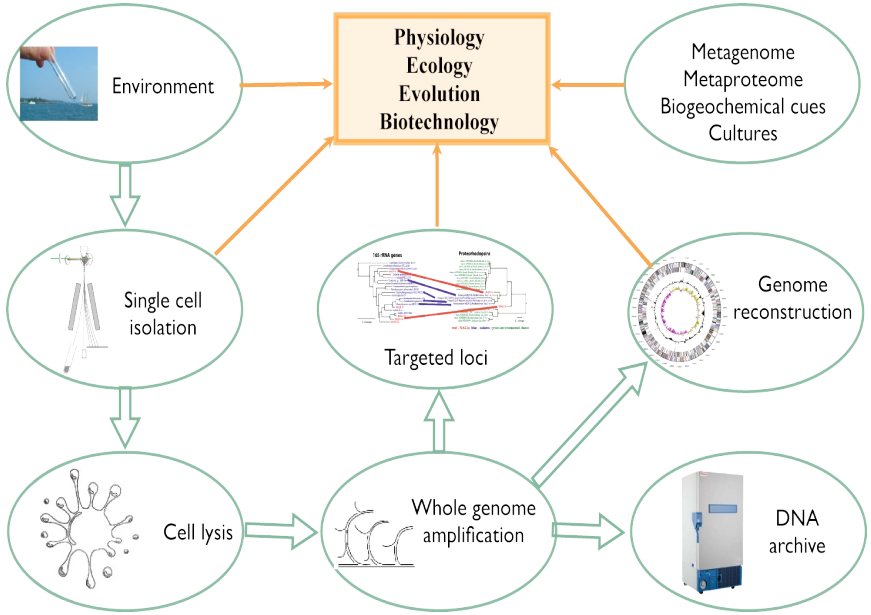
Contamination over-estimation by checkM



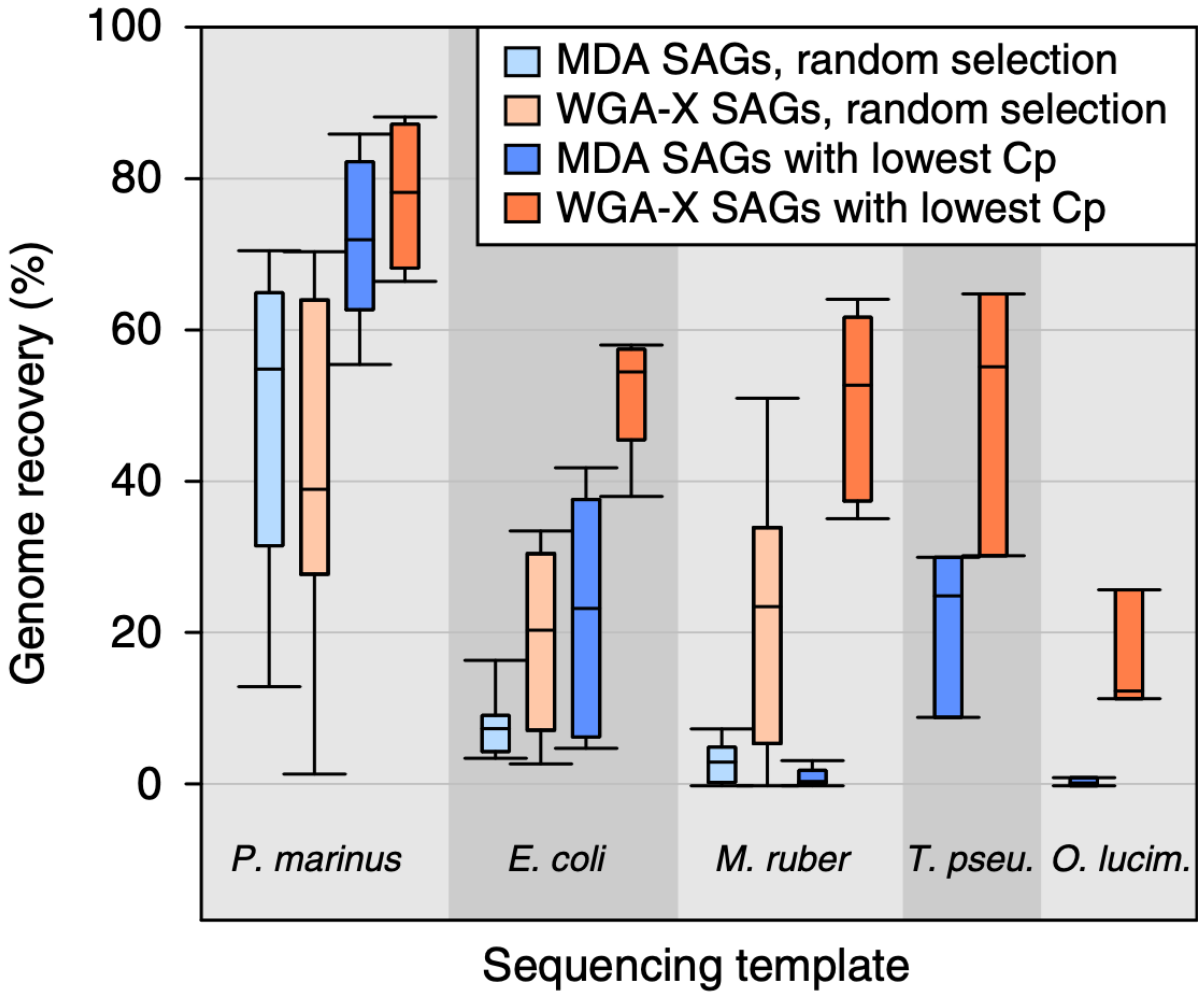
QA and QC at SCGC

SCGC strategy:

1. Validate the entire process
2. Minimize physical DNA contamination and damage
3. Maximize accuracy of the computational workflow
4. Generate data outputs for manual QC

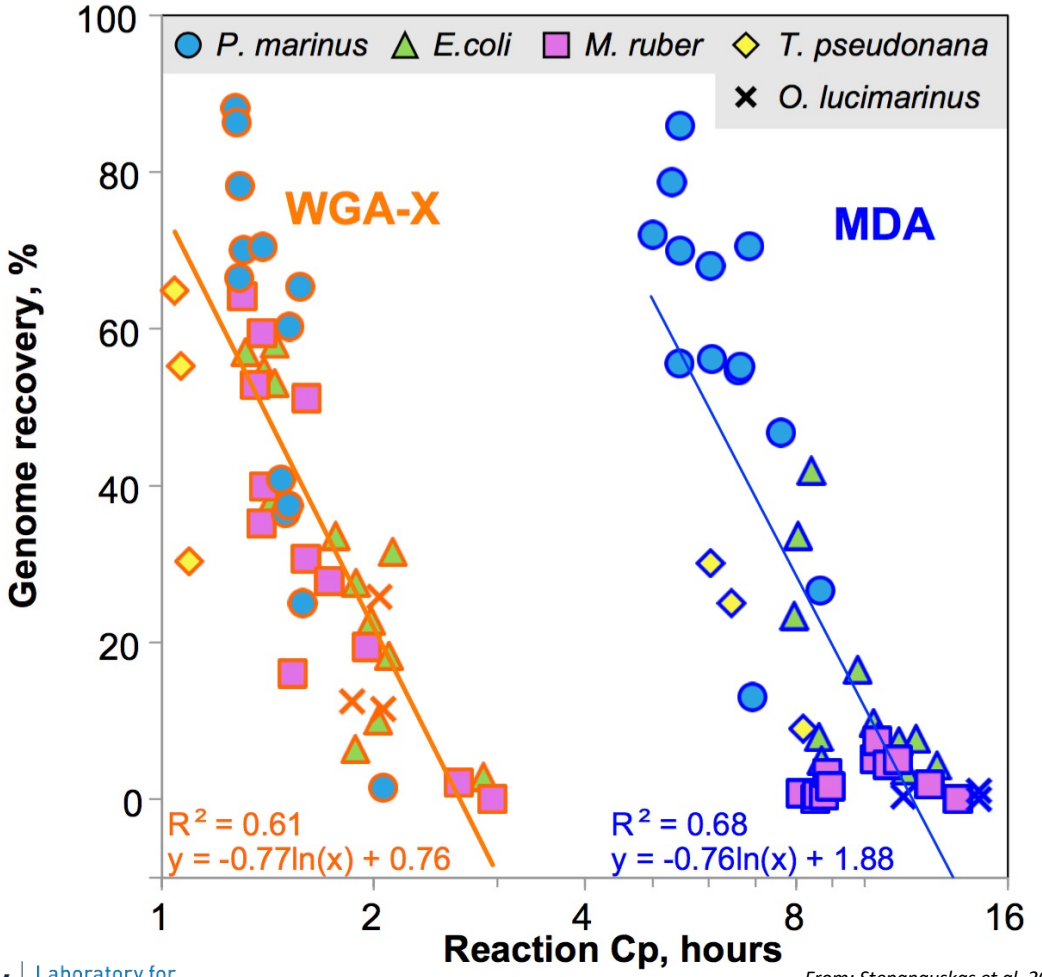


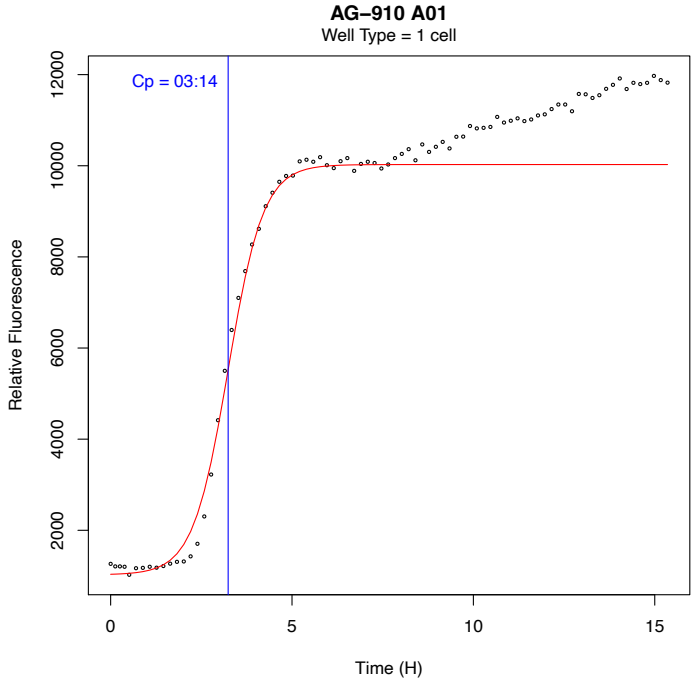
Validation of entire SCGC workflow



From: Stepanauskas et al. 2017

Validation of entire SCGC workflow

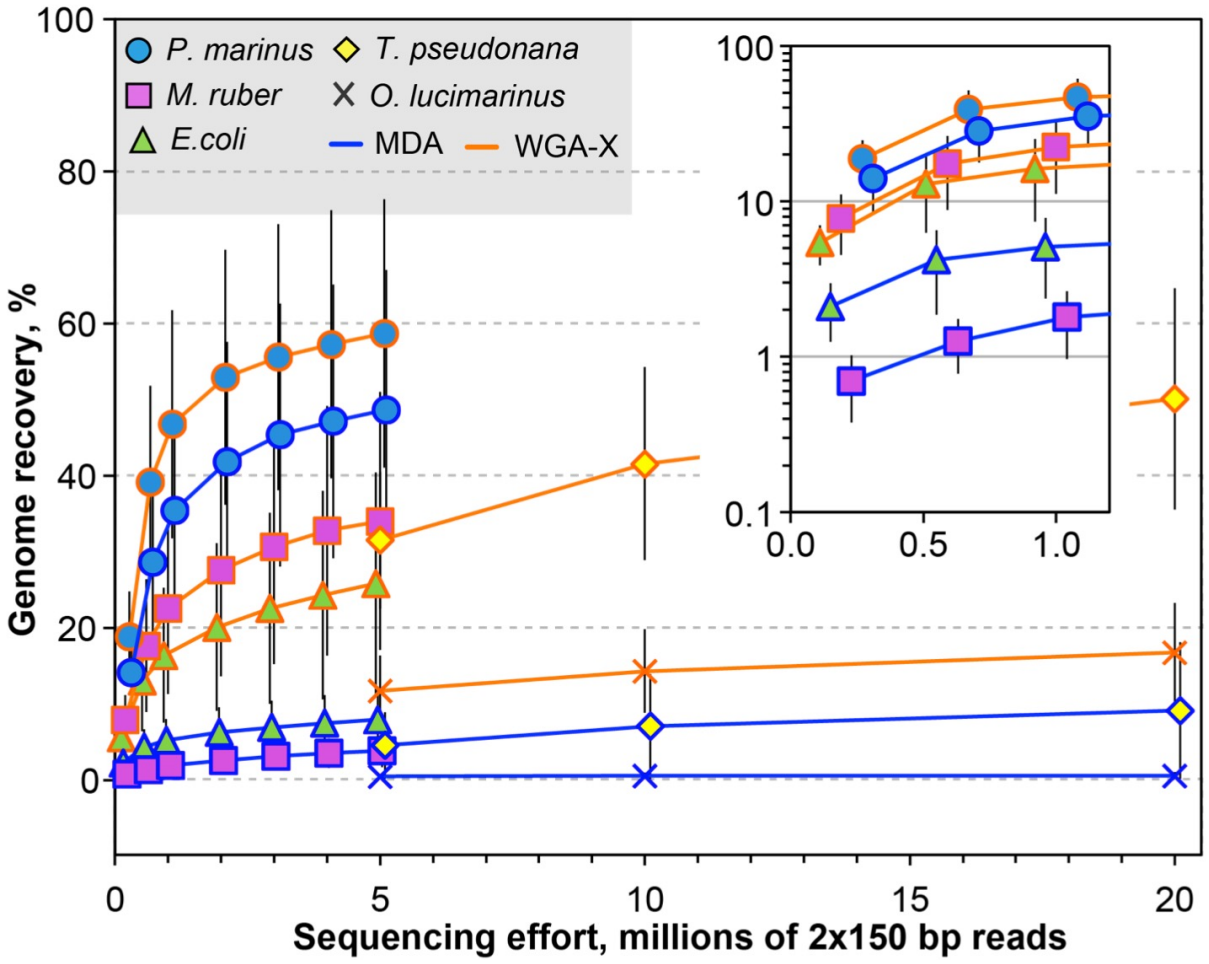




From: Stepanauskas et al. 2017

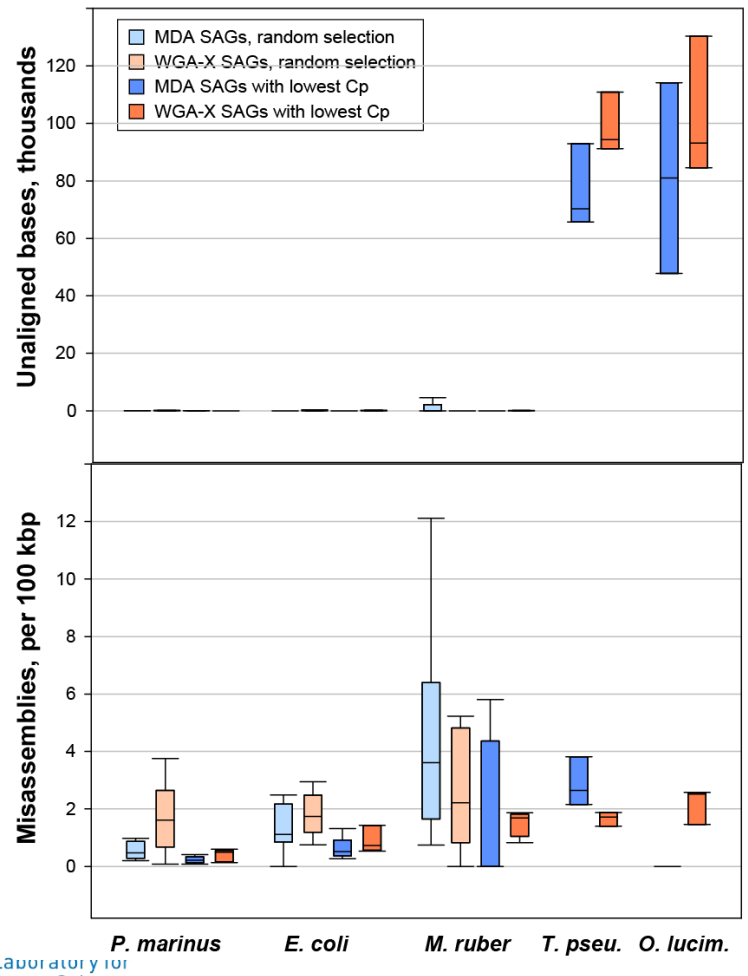
Bigelow | Laboratory for Ocean Sciences

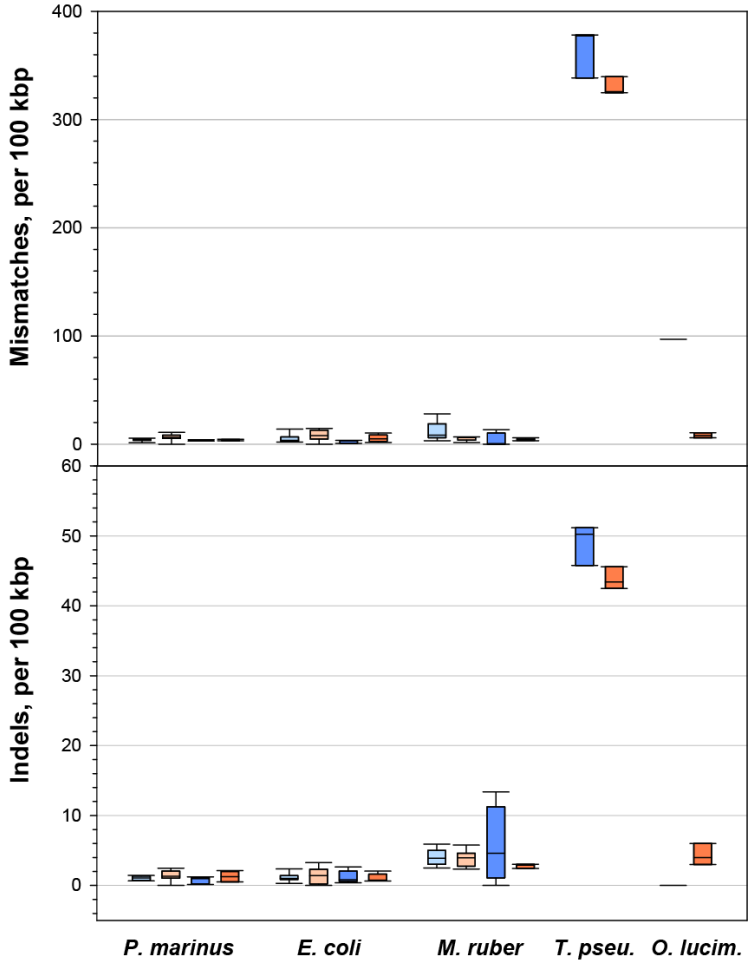
Relationship between sequencing depth and genome recovery



From: Stepanauskas et al. 2017

Validation of entire SCGC workflow



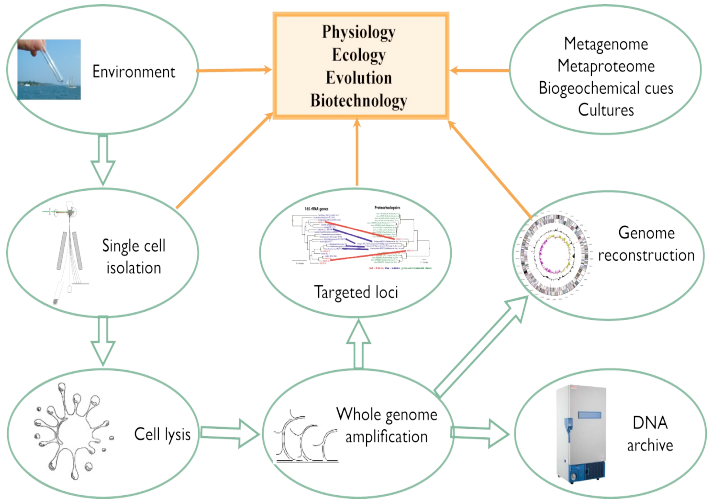


From: Stepanauskas et al. 2017

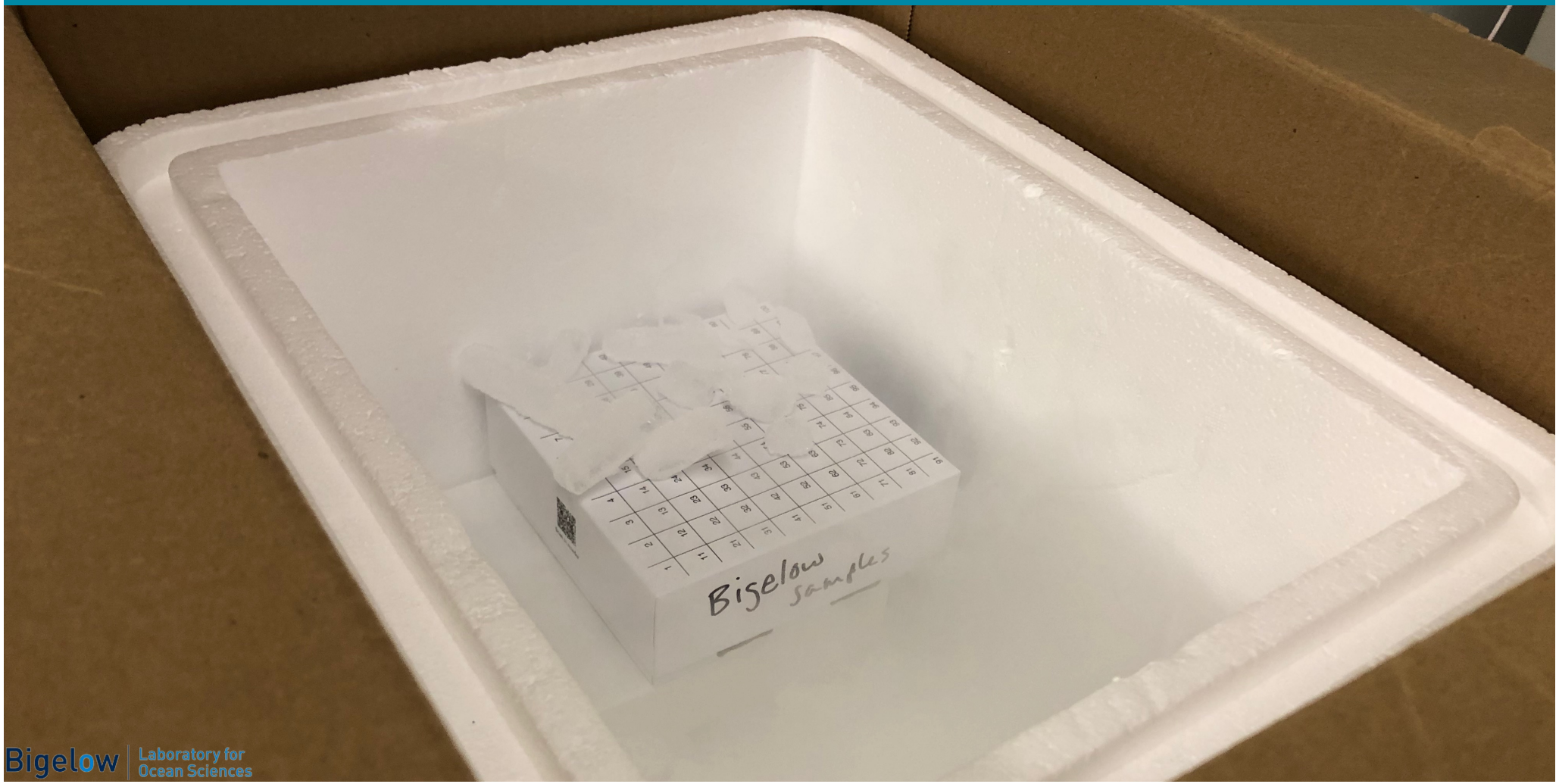
QA and QC at SCGC

SCGC strategy:

1. Validate the entire process
2. Minimize physical DNA contamination and damage
3. Maximize accuracy of the computational workflow
4. Generate data outputs for manual QC

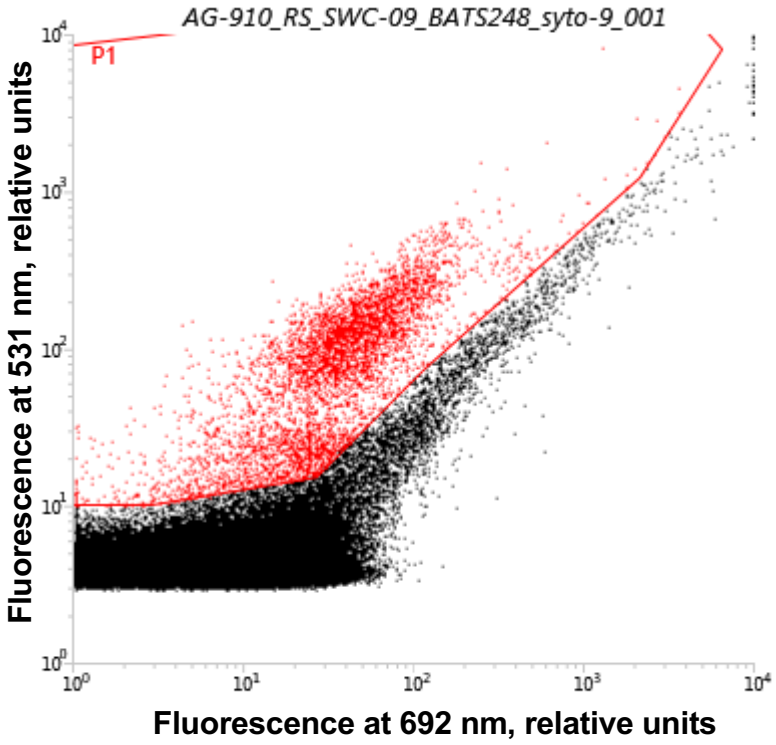


Pre-WGS QA: Field sample handling



Pre-WGS QA: Cleanroom environment

Pre-WGS QA: FACS gate



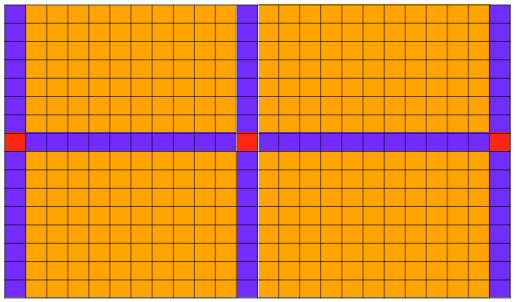
Pre-WGS QC: Whole genome amplification kinetics

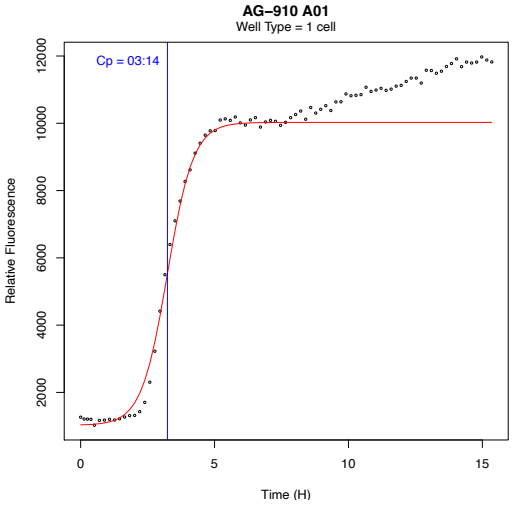
384-well plate layout

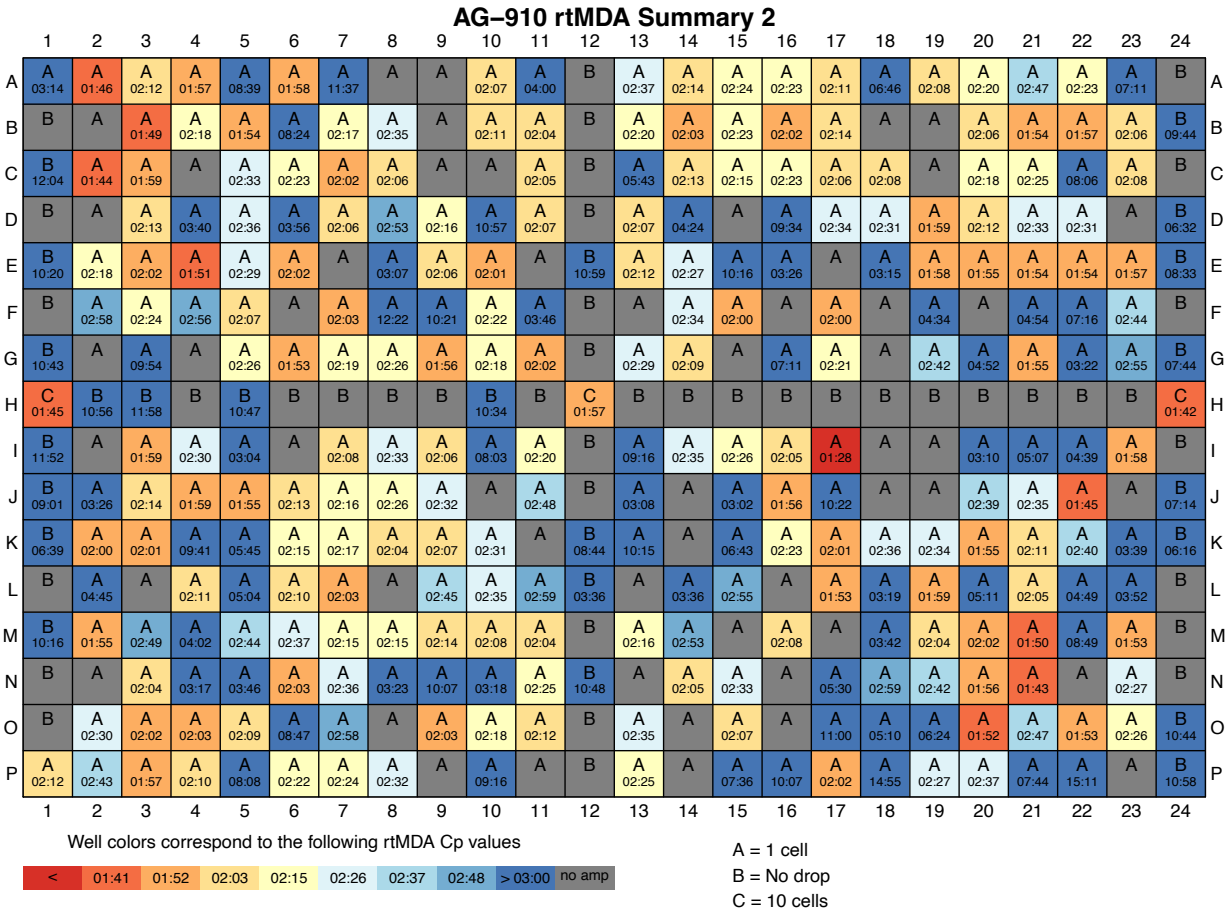
Negative controls: empty wells

Positive controls: 10 cells per well

Wells containing single cells







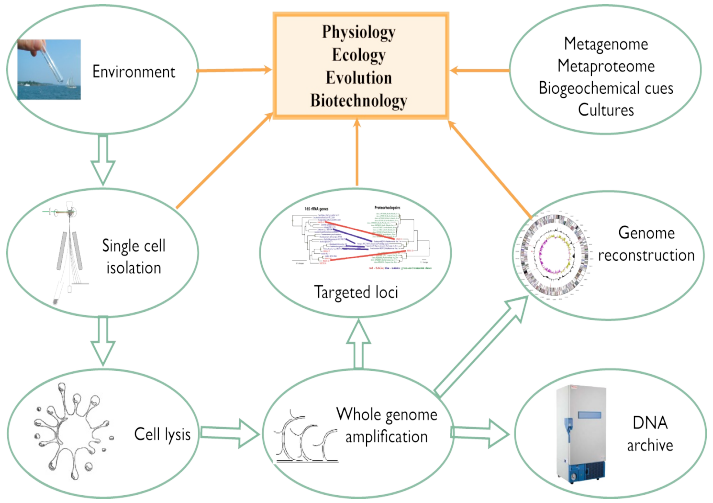
Contamination prevention and removal in SCGC's workflow

Measure	Type
Decontamination of instruments and reagents	Lab
Cleanroom for cell sorting and DNA amplification	Lab
Construction of a reagent contaminant database	Lab + Comp
Removal of reagent contaminants from reads and assemblies	Computation
Removal of contig ends, <2 kbp contigs, <20 kbp assemblies	Computation
Provision of tools for manual curation	Computation

QA and QC at SCGC

SCGC strategy:

1. Validate the entire process
2. Minimize physical DNA contamination and damage
3. Maximize accuracy of the computational workflow
4. Generate data outputs for manual QC



Tools for SCGC SAG assembly manual curation

Clue	Tool
Taxonomic assignment	Silva, (GTDBtk)
Extra copies of conserved single-copy genes	checkM
K-mer heterogeneity	Tetramer PCA
Heterogeneity of closest homologs	BLAST, (GUNC)
Annotation anomalies	Prokka, (DRAM)

Tools for SAG curation: Taxonomic assignments

/ storage / data /

Name

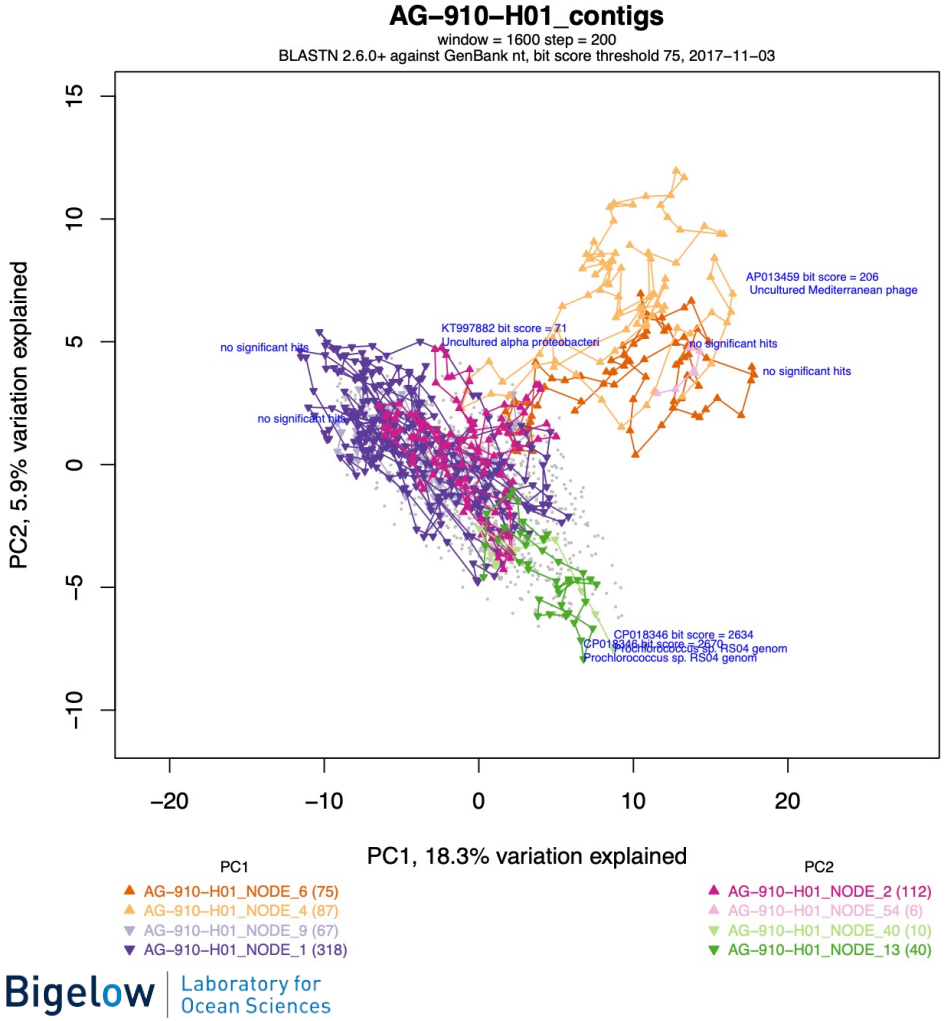
- contigs
- metagenomes
- sag_deliverables
- AG-910_assembly_stats_updated.csv
- AG-910_gorg-tropics_CelltableS2.xlsx
- AH-141_assembly_stats.xlsx
- AH-296_assembly_stats.xlsx
- bioinformatics_course_SAG_data.xlsx

Sample_ID	well_type	wga_cp	SSU_classification_1
AG-910-A16	1 cell	2.38	k__Bacteria;p__Actinobacteria;c__Acidimicrobiia;o__Acidimicrobiales;f__OM1_clade;g__Candidatus_Actinomarina;s__?
AG-910-L02	1 cell	4.75	k__Bacteria;p__Actinobacteria;c__Acidimicrobiia;o__Acidimicrobiales;f__OM1_clade;g__Candidatus_Actinomarina;s__?
AG-910-M10	1 cell	2.15	k__Bacteria;p__Actinobacteria;c__Acidimicrobiia;o__Acidimicrobiales;f__OM1_clade;g__Candidatus_Actinomarina;s__?
AG-910-D03	1 cell	2.23	k__Bacteria;p__Actinobacteria;c__Acidimicrobiia;o__Acidimicrobiales;f__OM1_clade;g__Candidatus_Actinomarina;s__?
AG-910-I07	1 cell	2.14	k__Bacteria;p__Bacteroidetes;c__Flavobacteriia;o__Flavobacteriales;f__Flavobacteriaceae;g__NS4_marine_group;s__?
AG-910-M02	1 cell	1.92	k__Bacteria;p__Bacteroidetes;c__Flavobacteriia;o__Flavobacteriales;f__Flavobacteriaceae;g__NS4_marine_group;s__?
AG-910-P04	1 cell	2.18	k__Bacteria;p__Bacteroidetes;c__Flavobacteriia;o__Flavobacteriales;f__Flavobacteriaceae;g__NS5_marine_group;s__?
AG-910-E10	1 cell	2.02	k__Bacteria;p__Cyanobacteria;c__Oxyphotobacteria;o__Synechococcales;f__Synechococcaceae;g__Prochlorococcus;s__?
AG-910-J04	1 cell	1.99	k__Bacteria;p__Cyanobacteria;c__Oxyphotobacteria;o__Synechococcales;f__Synechococcaceae;g__Prochlorococcus;s__?
AG-910-B21	1 cell	1.9	k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhodobacterales;f__Rhodobacteraceae;g__?;s__?
AG-910-B07	1 cell	2.3	k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhodospirillales;f__Rhodospirillaceae;g__AEGEAN-169_marine_group;s__?
AG-910-G06	1 cell	1.89	k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhodospirillales;f__Rhodospirillaceae;g__AEGEAN-169_marine_group;s__?
AG-910-G09	1 cell	1.94	k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhodospirillales;f__Rhodospirillaceae;g__AEGEAN-169_marine_group;s__?
AG-910-J16	1 cell	1.94	k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhodospirillales;f__Rhodospirillaceae;g__AEGEAN-169_marine_group;s__?
AG-910-K03	1 cell	2.03	k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhodospirillales;f__Rhodospirillaceae;g__AEGEAN-169_marine_group;s__?
AG-910-K06	1 cell	2.26	k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhodospirillales;f__Rhodospirillaceae;g__AEGEAN-169_marine_group;s__?
AG-910-N14	1 cell	2.09	k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhodospirillales;f__Rhodospirillaceae;g__AEGEAN-169_marine_group;s__?
AG-910-F04	1 cell	2.94	k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhodospirillales;f__Rhodospirillaceae;g__AEGEAN-169_marine_group;s__?
AG-910-E22	1 cell	1.91	k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rickettsiales;f__SAR116_clade;g__?;s__?
AG-910-D22	1 cell	2.52	k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rickettsiales;f__SAR116_clade;g__?;s__?
AG-910-M21	1 cell	1.85	k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__SAR11_clade;f__?;g__?;s__?
AG-910-C06	1 cell	2.38	k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__SAR11_clade;f__?;g__?;s__?
AG-910-C16	1 cell	2.38	k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__SAR11_clade;f__?;g__?;s__?
AG-910-D17	1 cell	2.57	k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__SAR11_clade;f__?;g__?;s__?
AG-910-E13	1 cell	2.21	k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__SAR11_clade;f__?;g__?;s__?
AG-910-M20	1 cell	2.05	k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__SAR11_clade;f__?;g__?;s__?
AG-910-N08	1 cell	3.39	k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__SAR11_clade;f__?;g__?;s__?
AG-910-N20	1 cell	1.94	k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__SAR11_clade;f__?;g__?;s__?
AG-910-E06	1 cell	2.04	k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__SAR11_clade;f__Surface_1;g__?;s__?

Tools for SAG curation: checkM

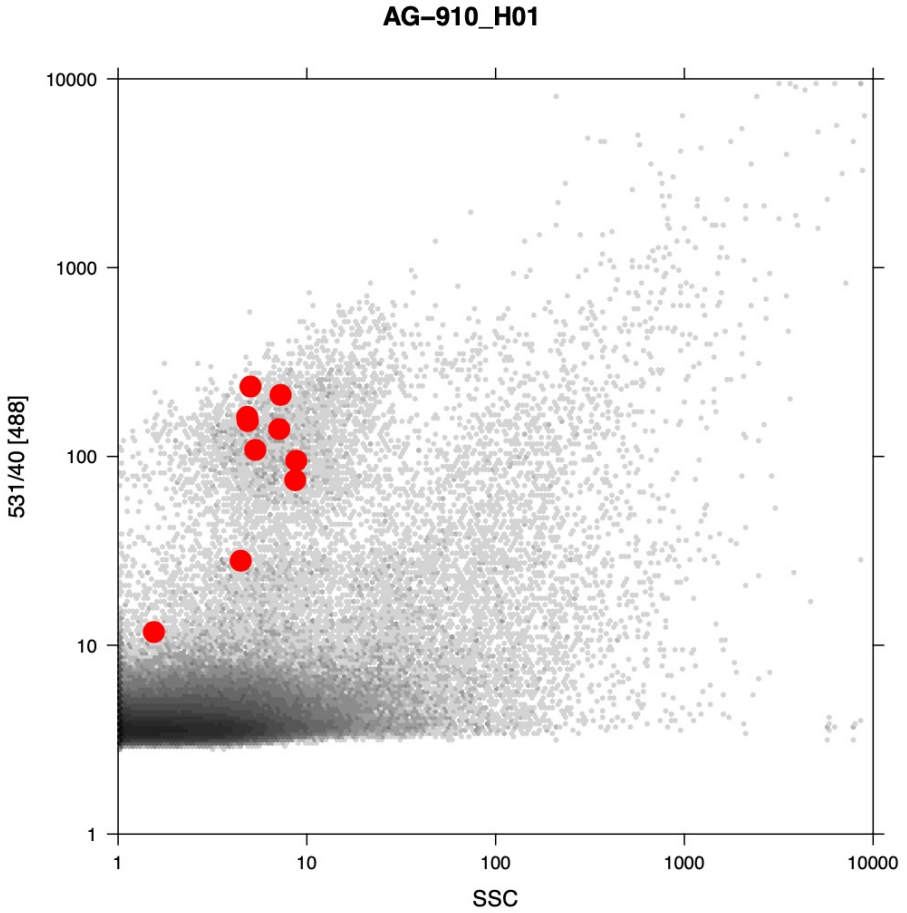
Sample_ID	well_type	wga_cp	final_assembly_length	max_contig_length	gc_content	checkM_estimated_completeness	number_multi_copy
AG-910-H24	10 cells	1.72	1,703,043	115,921	32	93	28
AG-910-N04	1 cell	3.29	214,674	27,694	29	22	2
AG-910-H12	10 cells	1.96	1,247,299	82,663	33	66	2
AG-910-K03	1 cell	2.03	1,067,614	103,436	30	56	1
AG-910-C18	1 cell	2.15	979,531	268,683	30	80	1
AG-910-I05	1 cell	3.08	83,974	27,163	32	4	1
AG-910-A16	1 cell	2.38	649,299	87,176	33	45	0
AG-910-L02	1 cell	4.75	134,319	41,293	32	14	0
AG-910-M10	1 cell	2.15	757,654	179,180	33	54	0
AG-910-D03	1 cell	2.23	127,585	56,880	29	11	0
AG-910-I07	1 cell	2.14	956,113	137,377	31	55	0
AG-910-M02	1 cell	1.92	1,128,449	191,928	31	73	0
AG-910-P04	1 cell	2.18	1,088,902	84,609	31	49	0
AG-910-E10	1 cell	2.02	1,247,571	148,843	31	84	0
AG-910-J04	1 cell	1.99	1,196,796	116,970	31	68	0
AG-910-B21	1 cell	1.9	1,930,694	375,705	31	75	0
AG-910-B07	1 cell	2.3	751,372	109,892	31	61	0
AG-910-G06	1 cell	1.89	1,155,577	330,630	31	70	0

Tetramer PCA: Multiple cells



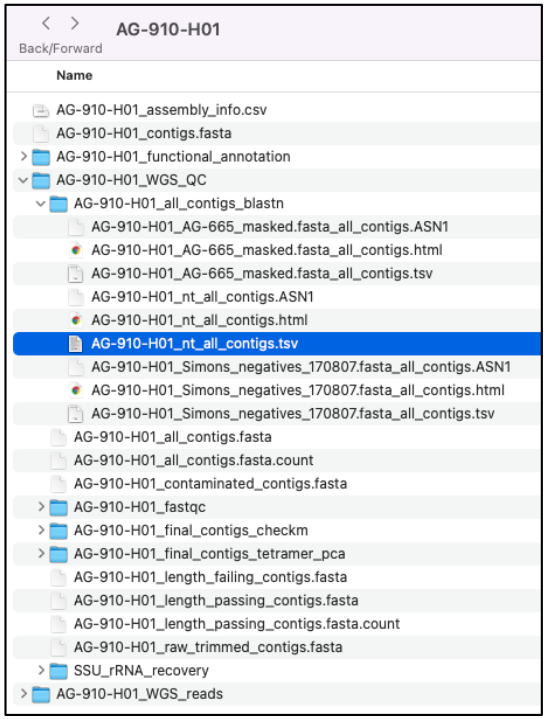
Bigelow | Laboratory for Ocean Sciences

FACS: Particle light scatter and green fluorescence



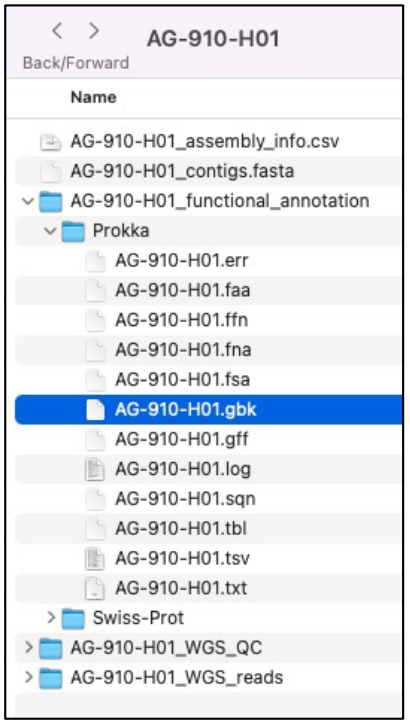
Assembly QC: BLASTN on contigs

Query Seq-id	Sub	Percentage of identical matches	Alignment length	gth	All Subject Title(s)
AG-910-H01_NODE_4	gi|5	92	1182	82	Uncultured virus clone ctg_DTF_polA_1102 putative DNA polymerase A gene, partial cds
AG-910-H01_NODE_4	gi|5	74	574	25	Uncultured Mediterranean phage uvMED DNA, complete genome, group G17, isolate: uvMED-CGR-U-MedDCM-OCT-S39-C21
AG-910-H01_NODE_4	gi|1	75	337	97	Uncultured virus clone vSAG-41-A4-2 genomic sequence
AG-910-H01_NODE_4	gi|1	75	334	97	Uncultured virus clone vSAG-41-A4-2 genomic sequence
AG-910-H01_NODE_4	gi|5	81	101	64	Uncultured Mediterranean phage uvMED DNA, complete genome, group G19, isolate: uvMED-CGR-U-MedDCM-OCT-S24-C60
AG-910-H01_NODE_4	gi|1	91	43	05	Pichia membranifaciens NRRL Y-2026 hypothetical protein partial mRNA



Assembly QC: Prokka annotation

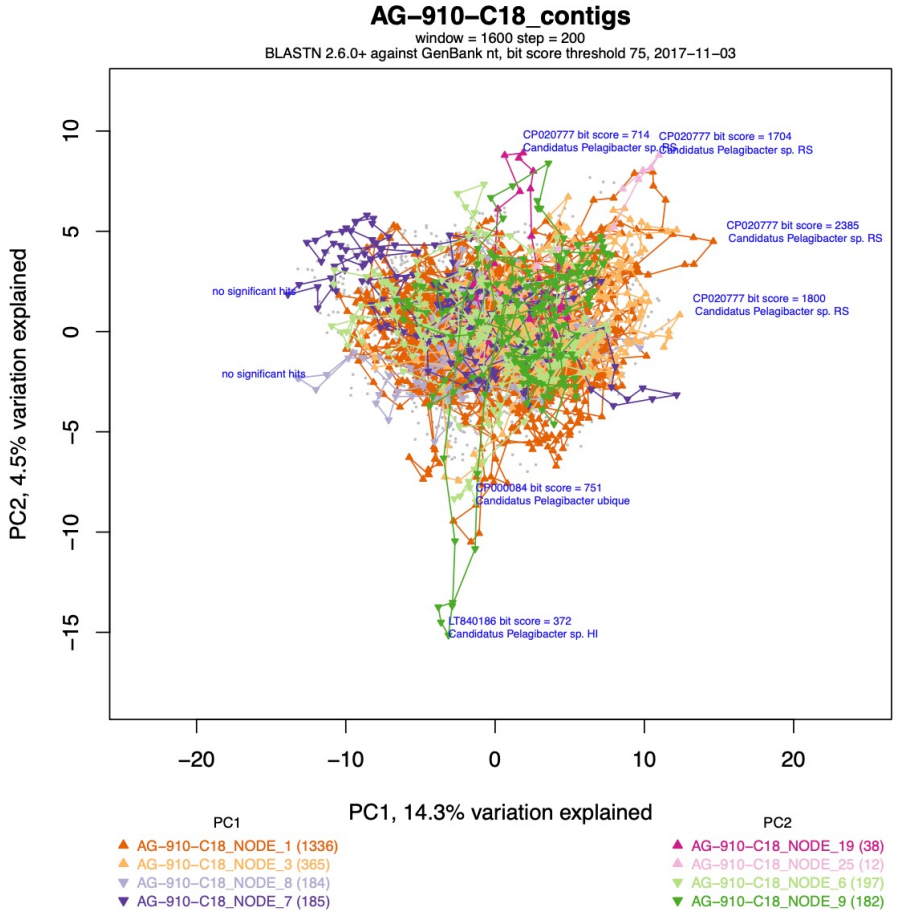
```
//
LOCUS       AG-910-H01_NODE_4      18923 bp    DNA     linear       03-NOV-2017
DEFINITION  Genus species strain strain.
ACCESSION
VERSION
KEYWORDS    .
SOURCE      Genus species
  ORGANISM  Genus species
            Unclassified.
COMMENT     Annotated using prokka 1.12 from
            https://github.com/tseemann/prokka.
FEATURES             Location/Qualifiers
     source          1..18923
                     /organism="Genus species"
                     /mol_type="genomic DNA"
                     /strain="strain"
     CDS             21..131
                     /locus_tag="AG-910-H01_00107"
                     /inference="ab initio prediction:Prodigal:2.6"
                     /codon_start=1
                     /transl_table=11
                     /product="hypothetical protein"
                     /translation="MSNKCKYCNGNCPNDSDYMCDGYAGDIDNLYENNEE"
     CDS             134..571
                     /locus_tag="AG-910-H01_00108"
                     /inference="ab initio prediction:Prodigal:2.6"
                     /codon_start=1
                     /transl_table=11
                     /product="hypothetical protein"
                     /translation="MNMDRLLESVKKHEGYRNKVYLDTLGKRTVGVGHLCVEDFWEDN
                     KEYEEKFLMTILEHDLQTAIKGAKELMEDHGCADIDEQAEEILIEMVFQLGKNGVSKF
                     KNMWKALAEKNYIGASYEMLDSRWAKQTPNRAKSMAKTMKEIT"
     CDS             568..753
                     /locus_tag="AG-910-H01_00109"
                     /inference="ab initio prediction:Prodigal:2.6"
                     /codon_start=1
                     /transl_table=11
                     /product="hypothetical protein"
                     /translation="MIYKGQKLSDNLTKSQRDMFKYMIDRGDSIEEIVKIYTLDDKVN
                     QALREPDVEEKIEALKE"
```



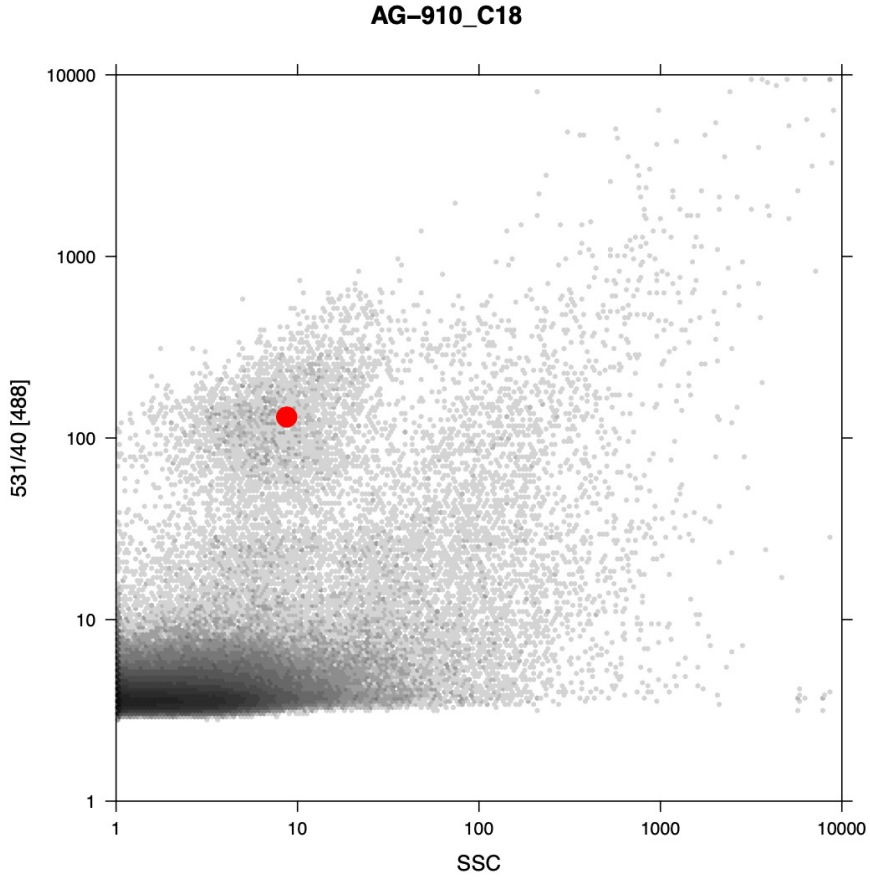
Tools for SAG curation: checkM

Sample_ID	well_type	wga_cp	final_assembly_length	max_contig_length	gc_content	checkM_estimated_completeness	number_multi_copy
AG-910-H24	10 cells	1.72	1,703,043	115,921	32	93	28
AG-910-N04	1 cell	3.29	214,674	27,694	29	22	2
AG-910-H12	10 cells	1.96	1,247,299	82,663	33	66	2
AG-910-K03	1 cell	2.03	1,067,614	103,436	30	56	1
AG-910-C18	1 cell	2.15	979,531	268,683	30	80	1
AG-910-I05	1 cell	3.08	83,974	27,163	32	4	1
AG-910-A16	1 cell	2.38	649,299	87,176	33	45	0
AG-910-L02	1 cell	4.75	134,319	41,293	32	14	0
AG-910-M10	1 cell	2.15	757,654	179,180	33	54	0
AG-910-D03	1 cell	2.23	127,585	56,880	29	11	0
AG-910-I07	1 cell	2.14	956,113	137,377	31	55	0
AG-910-M02	1 cell	1.92	1,128,449	191,928	31	73	0
AG-910-P04	1 cell	2.18	1,088,902	84,609	31	49	0
AG-910-E10	1 cell	2.02	1,247,571	148,843	31	84	0
AG-910-J04	1 cell	1.99	1,196,796	116,970	31	68	0
AG-910-B21	1 cell	1.9	1,930,694	375,705	31	75	0
AG-910-B07	1 cell	2.3	751,372	109,892	31	61	0
AG-910-G06	1 cell	1.89	1,155,577	330,630	31	70	0

Tetramer PCA: Clean SAG

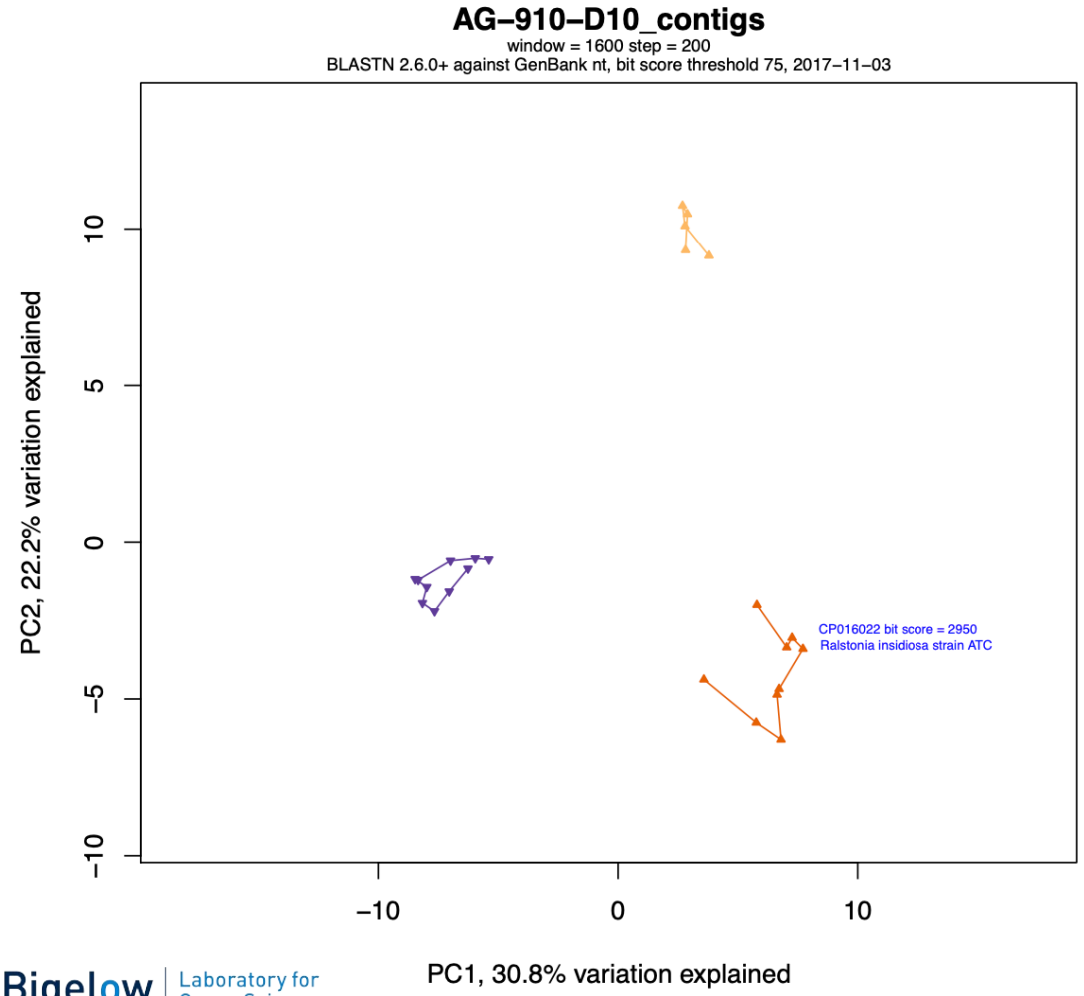


FACS: Particle light scatter and green fluorescence

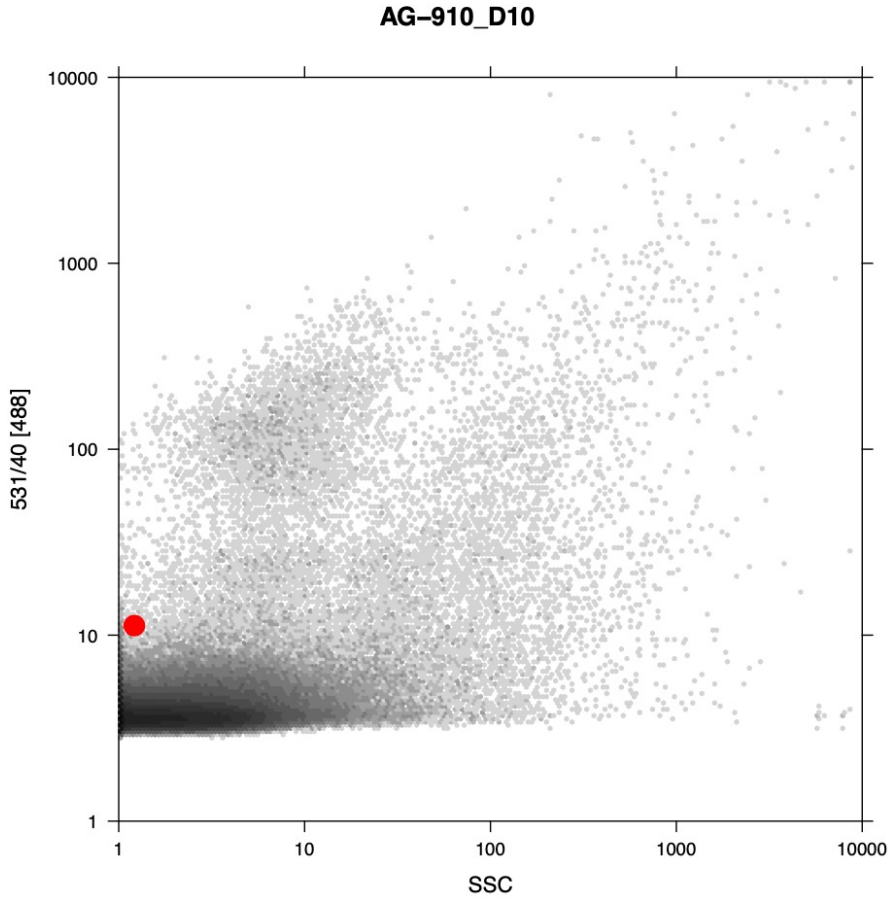


Bigelow | Laboratory for Ocean Sciences

Tetramer PCA: Reagent contaminants

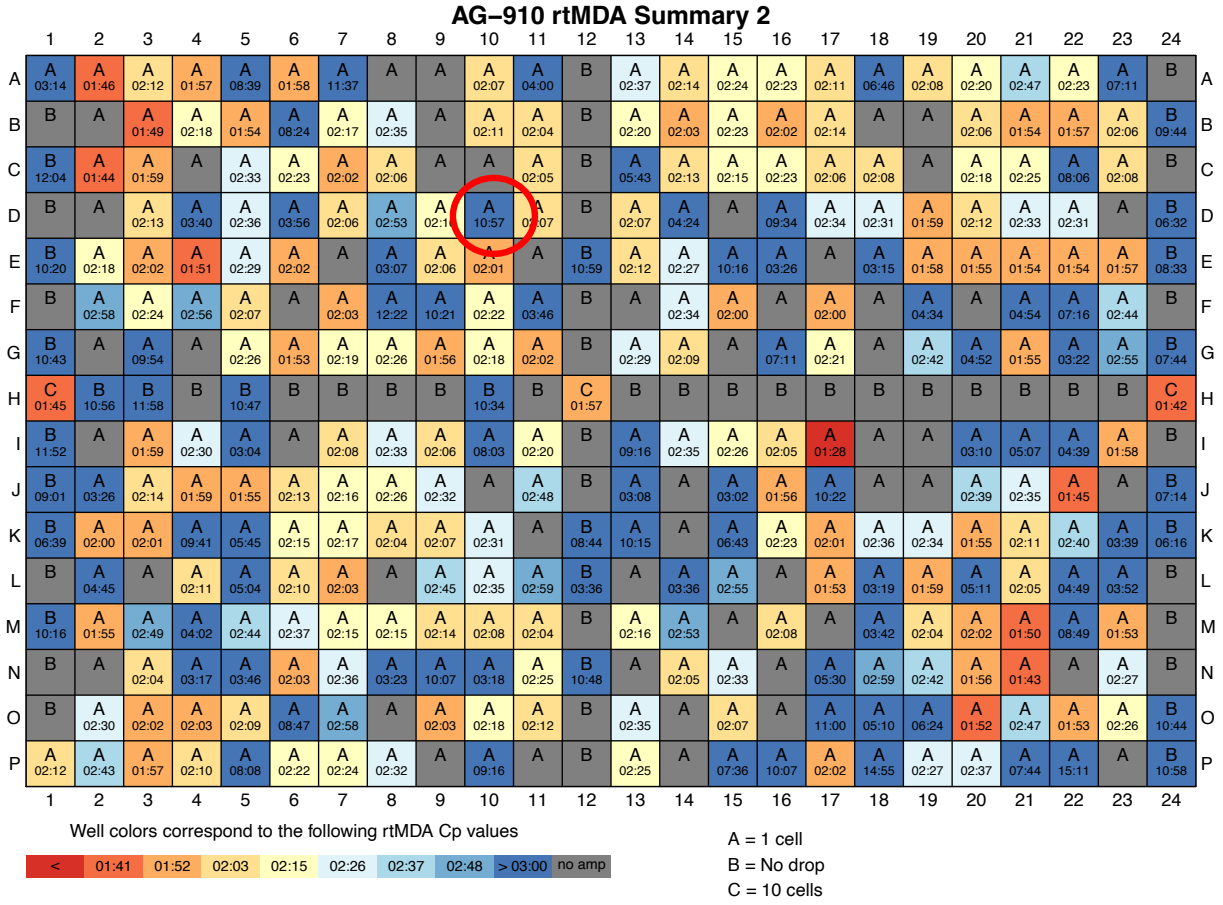


FACS: Particle light scatter and green fluorescence



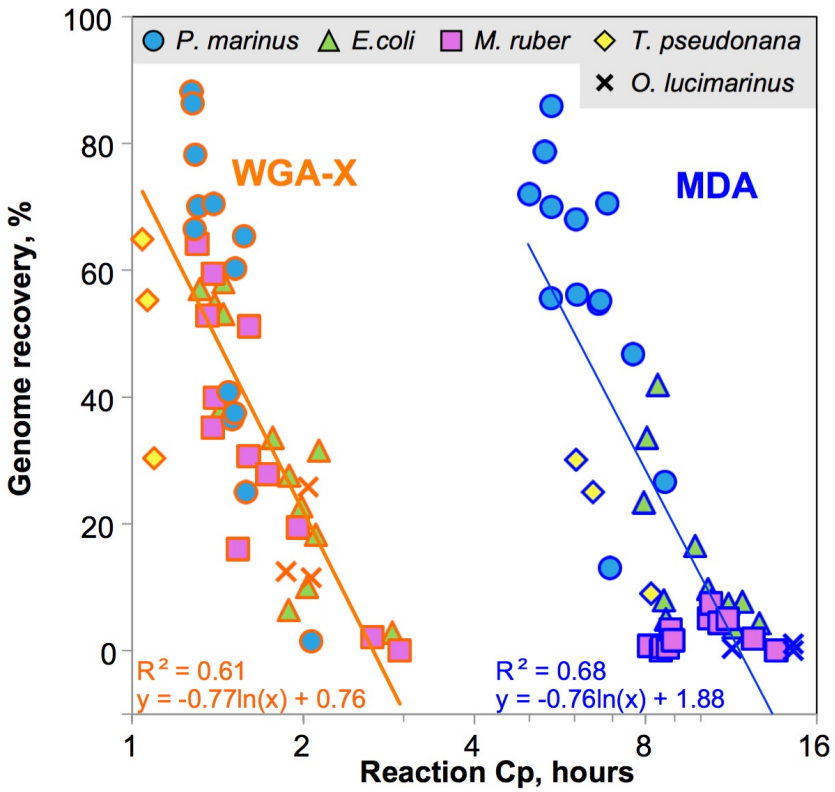
Bigelow | Laboratory for Ocean Sciences

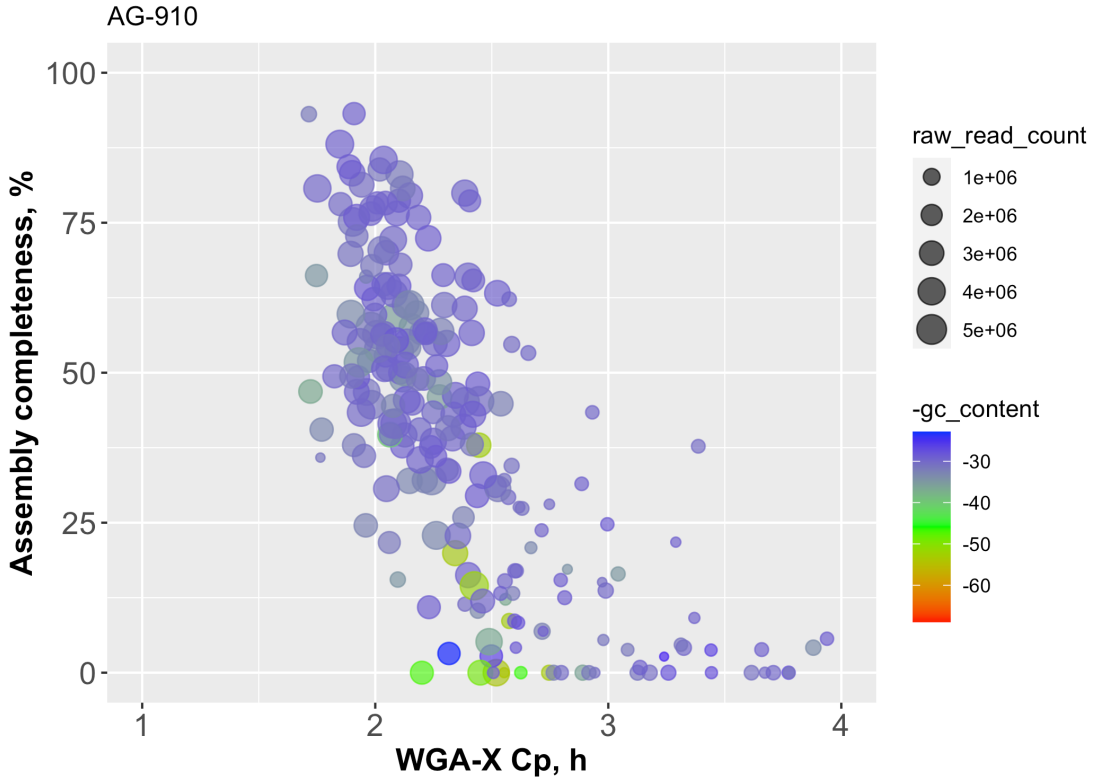
Pre-WGS QC: Whole genome amplification kinetics



Bigelow | Laboratory for Ocean Sciences

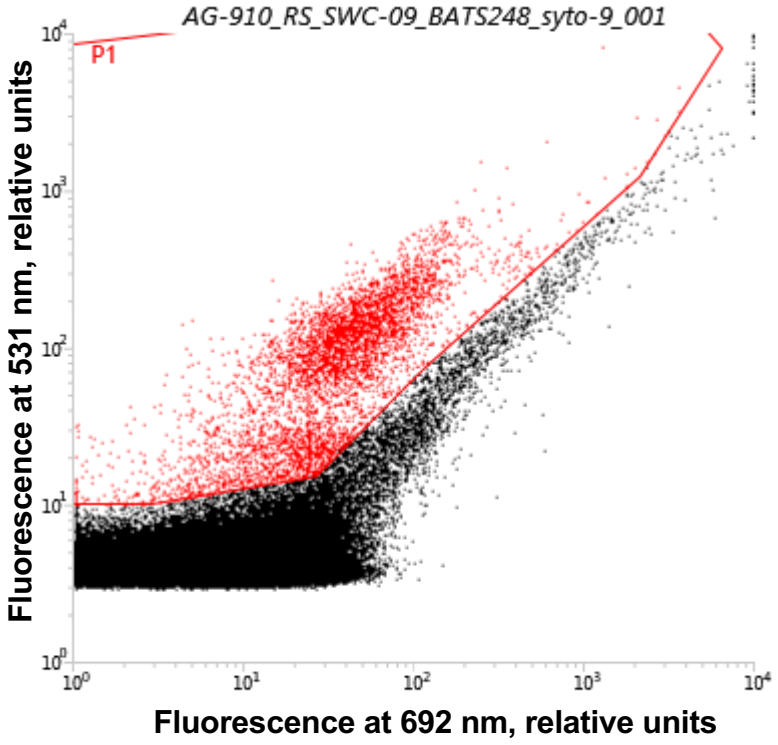
Multi-parameter visualization of AG-910 SAGs

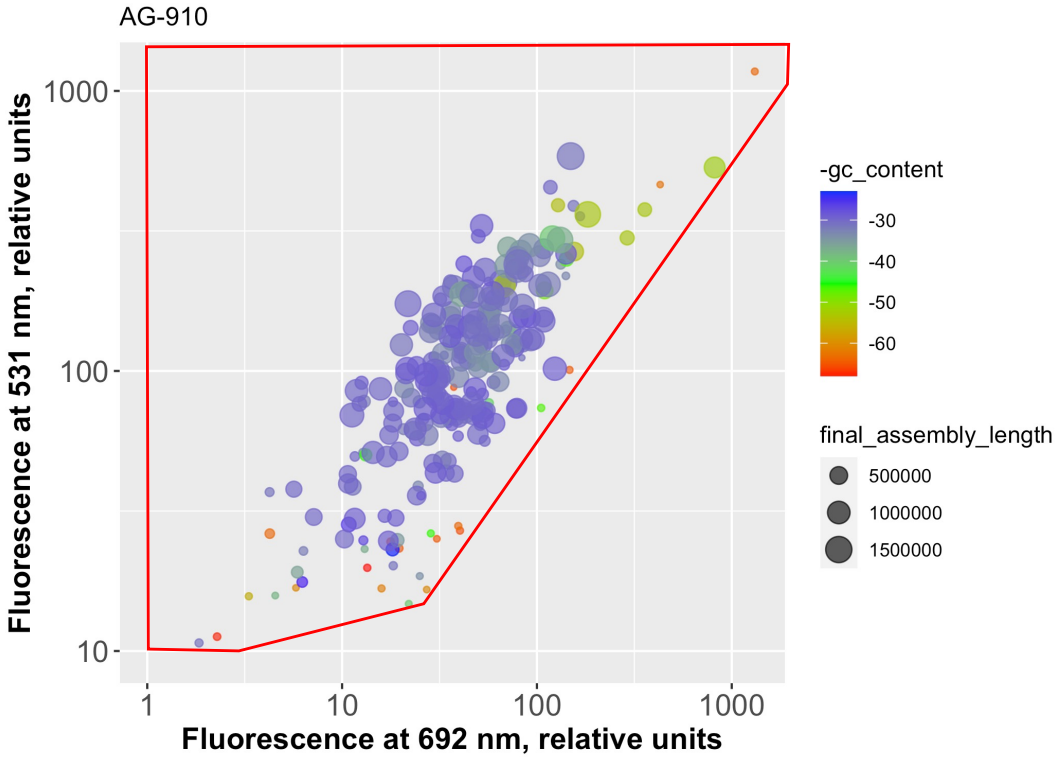




Bigelow | Laboratory for Ocean Sciences

Multi-parameter visualization of AG-910 SAGs

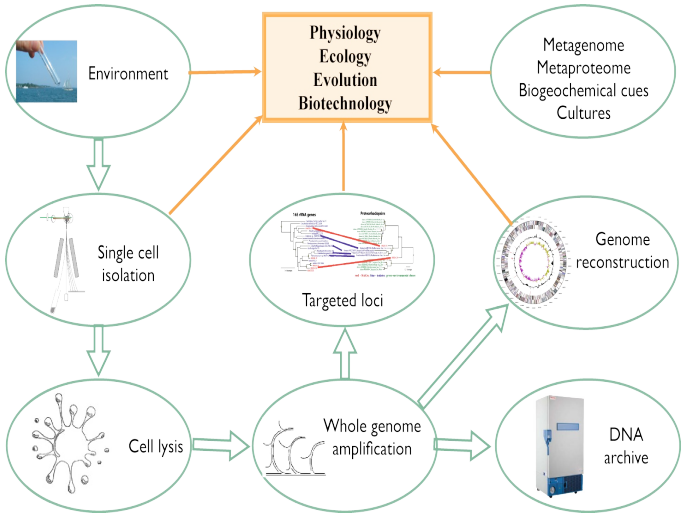




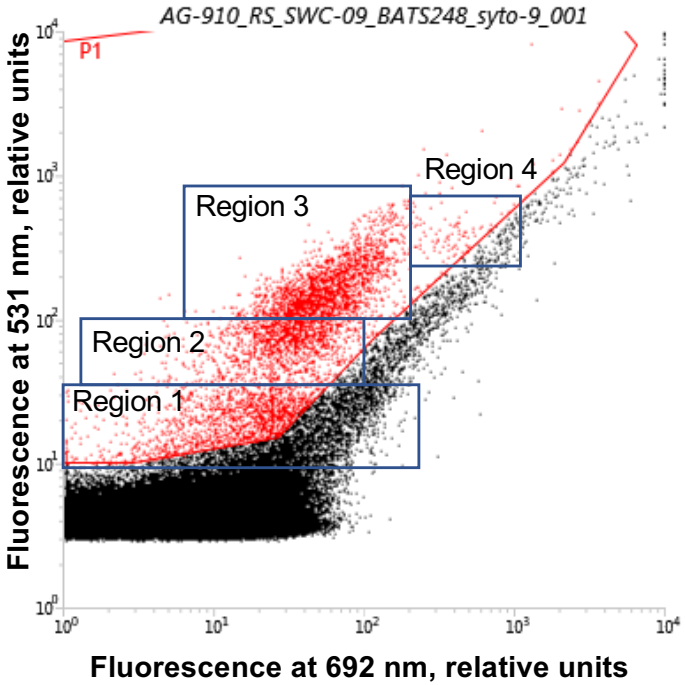
Bigelow | Laboratory for Ocean Sciences

Take-home message

1. Use genome QC metrics and thresholds that match your research goals.
2. Understand how the data were produced, field -> lab -> computation.
3. Be critical: no tool is perfect; run reality checks and use all available information.



SCGC data exploration exercise



What particles were in FACS regions 1-4?