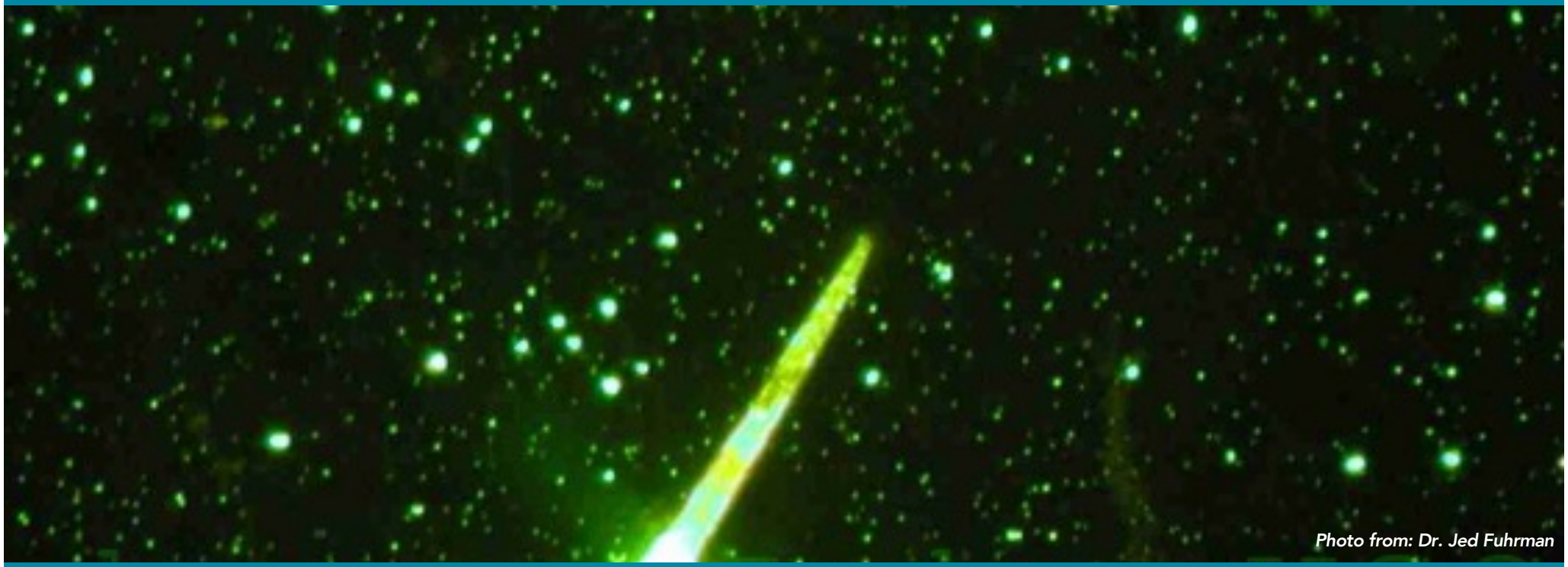
Bigelow | Laboratory for Ocean Sciences

Ramunas Stepanauskas



SCGC quality assessment and control

How do we define genome assembly quality?

nature biotechnology

OPEN

2017

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

Robert M Bowers[1], Nikos C Kyrpides[1], Ramunas Stepanauskas[2], Miranda Harmon-Smith[1], Devin Doud[1], T B K Reddy[1], Frederik Schulz[1], Jessica Jarett[1], Adam R Rivers[1,3], Emiley A Eloe-Fadrosh[1], Susannah G Tringe[1,4], Natalia N Ivanova[1], Alex Copeland[1], Alicia Clum[1], Eric D Becraft[2], Rex R Malmstrom[1], Bruce Birren[5], Mircea Podar[6], Peer Bork[7], George M Weinstock[8], George M Garrity[9], Jeremy A Dodsworth[10], Shibu Yooseph[11], Granger Sutton[12], Frank O Glöckner[13], Jack A Gilbert[14,15], William C Nelson[16], Steven J Hallam[17], Sean P Jungbluth[1,18], Thijs J G Ettema[19], Scott Tighe[20], Konstantinos T Konstantinidis[21], Wen-Tso Liu[22], Brett J Baker[23], Thomas Rattei[24], Jonathan A Eisen[25], Brian Hedlund[26,27], Katherine D McMahon[28,29], Noah Fierer[30,31], Rob Knight[32], Rob Finn[33], Guy Cochrane[33], Ilene Karsch-Mizrachi[34], Gene W Tyson[35], Christian Rinke[35] The Genome Standards Consortium[36], Alla Lapidus[37], Folker Meyer[14], Pelin Yilmaz[13], Donovan H Parks[35], A M Eren[38], Lynn Schriml[39], Jillian F Banfield[40], Philip Hugenholtz[35] & Tanja Woyke[1,4]

Bigelow | Laboratory Ocean Sciences

How do we define genome assembly quality?

Table 1 Genome reporting standards for SAGs and MAGs

Criterion	Description
Finished (SAG/MAG)	
Assembly quality[a]	Single contiguous sequence without gaps or ambiguities with a consensus error rate equivalent to Q50 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 (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%
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%

How do we define genome assembly quality?

Downloaded from genome.cshlp.org on January 18, 2016 - Published by Cold Spring Harbor Laboratory Press

Method

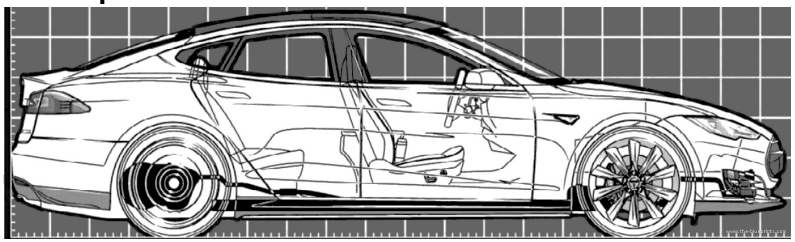
CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes

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

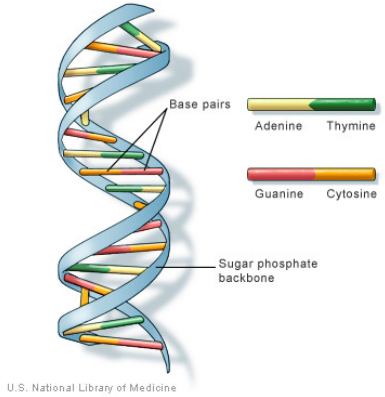
Genome: organism's blueprint

Blueprint of a car



From: the-blueprints.com

Blueprint (genome) of a microbe

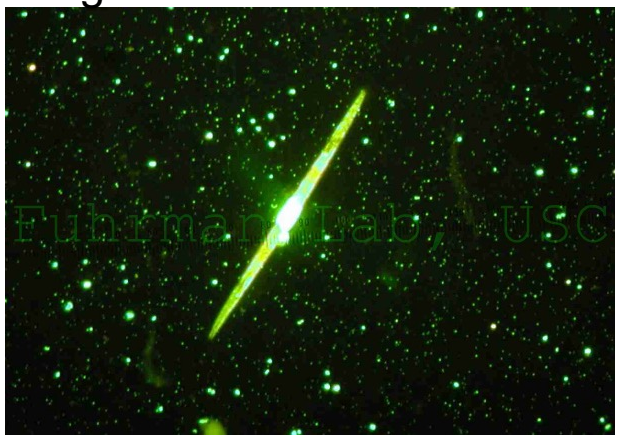


From: istockphoto.com

Image of a car

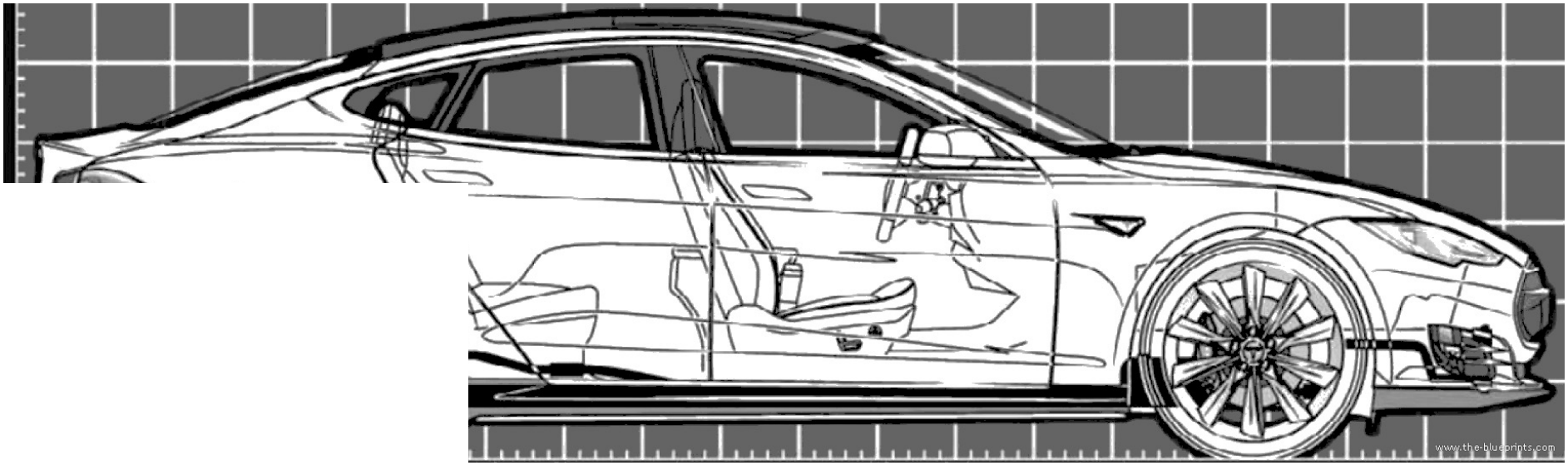
From: tautoblog.com

Image of marine microbes

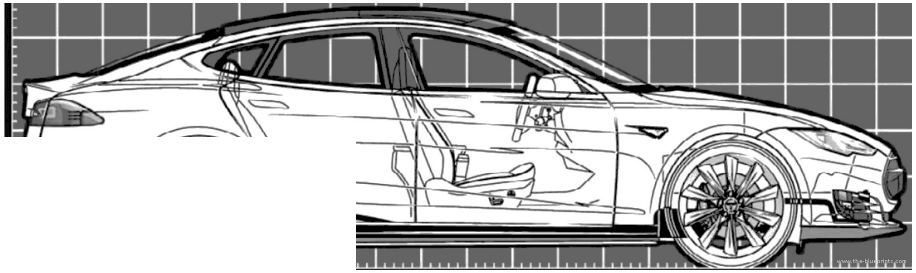


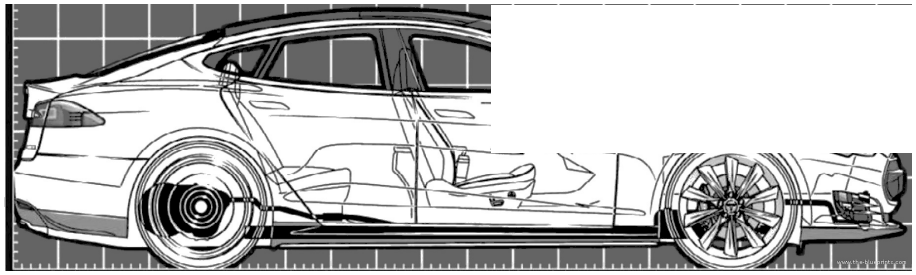
From: Jed Fuhrman

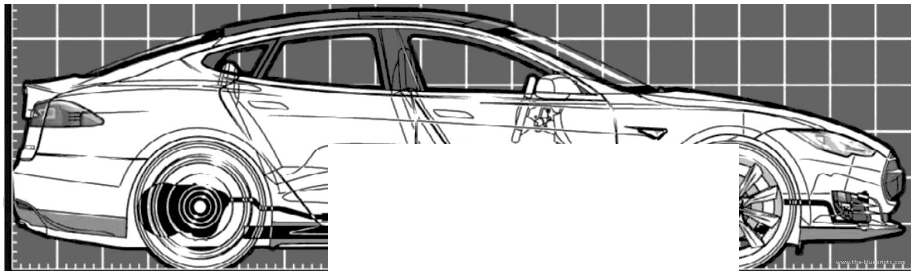
Blueprint incompleteness



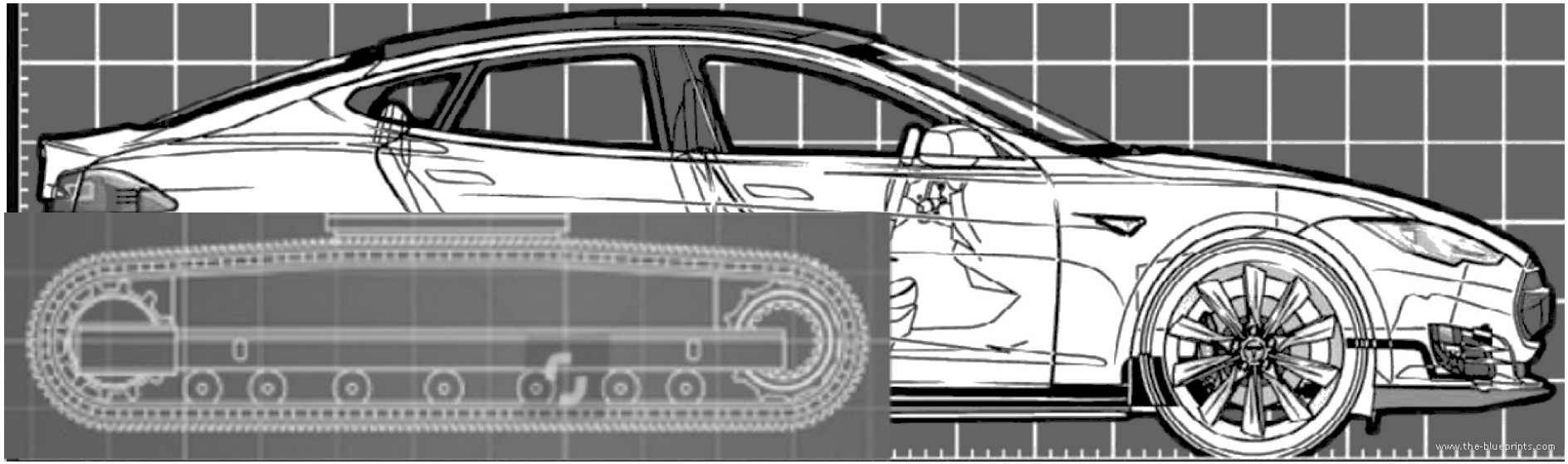
Compensation for incompleteness of individual blueprints







Blueprint chimerism (contamination)



Contamination under-estimation by checkM

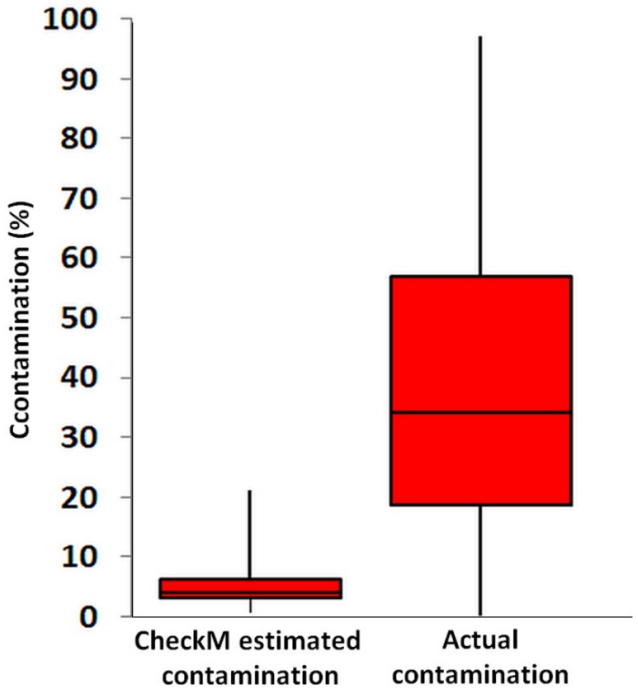


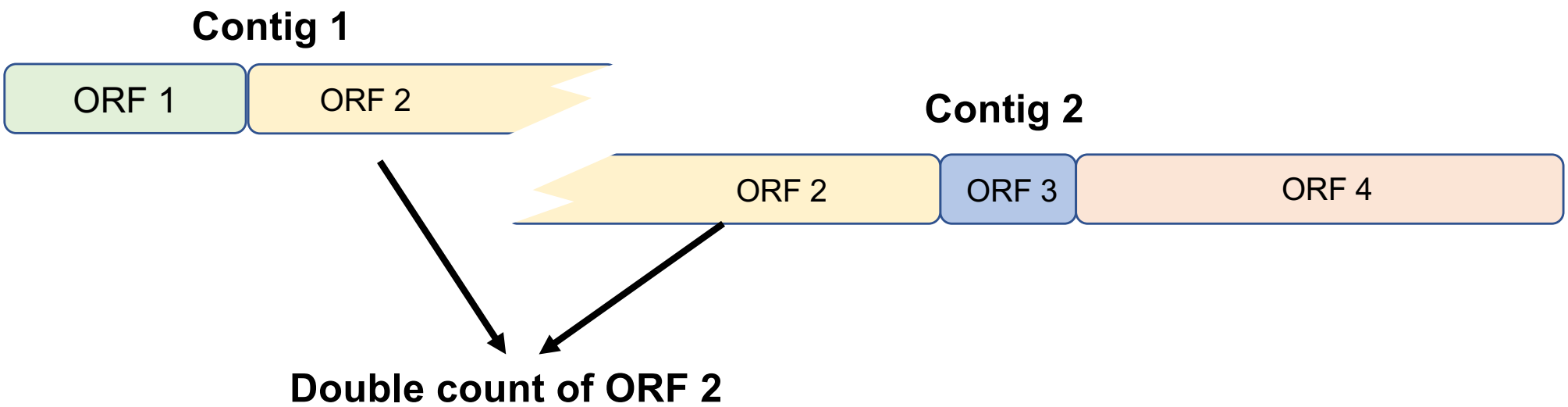
FIGURE 6 | Contamination predicted by CheckM software from all pairwise SAG combinations across the class-level lineages Rokumicrobia and Infratellusbacteria (15 Rokumicrobia and 4 Infratellusbacteria SAGs; 60 combinations total), compared to actual contamination calculated from all artificially combined 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 .

Contamination over-estimation by checkM



Other tools for contamination detection

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*]

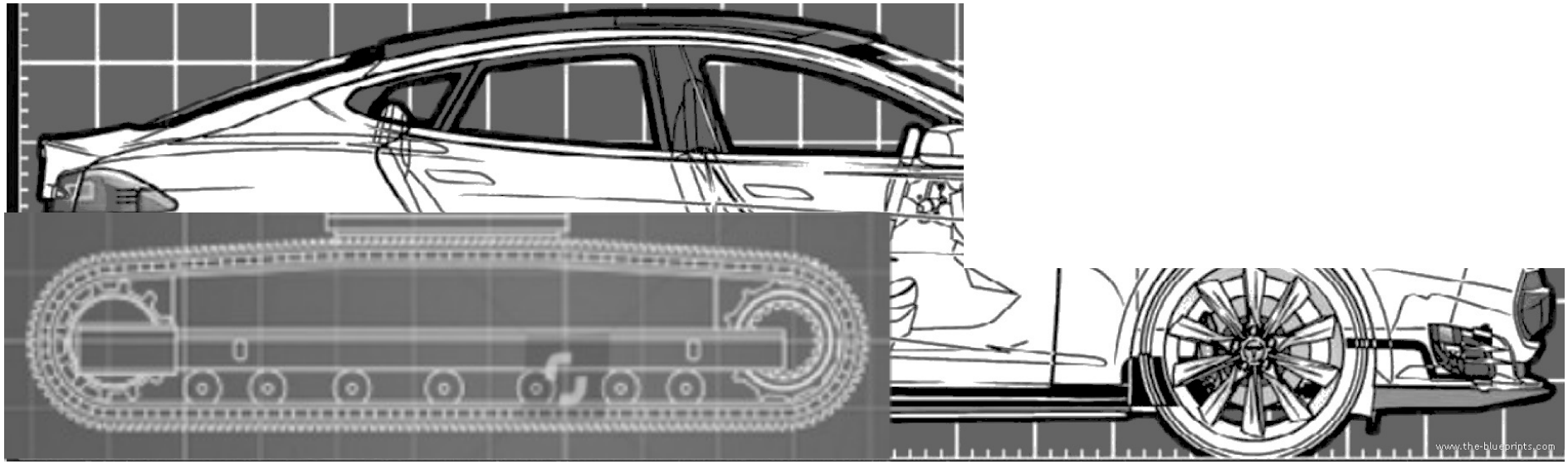
Nucleic Acids Research, 2022 **1**
https://doi.org/10.1093/nar/gkac294

How clear is our current view on microbial dark matter? (Re-)assessing public MAG & SAG datasets with MDMcleaner

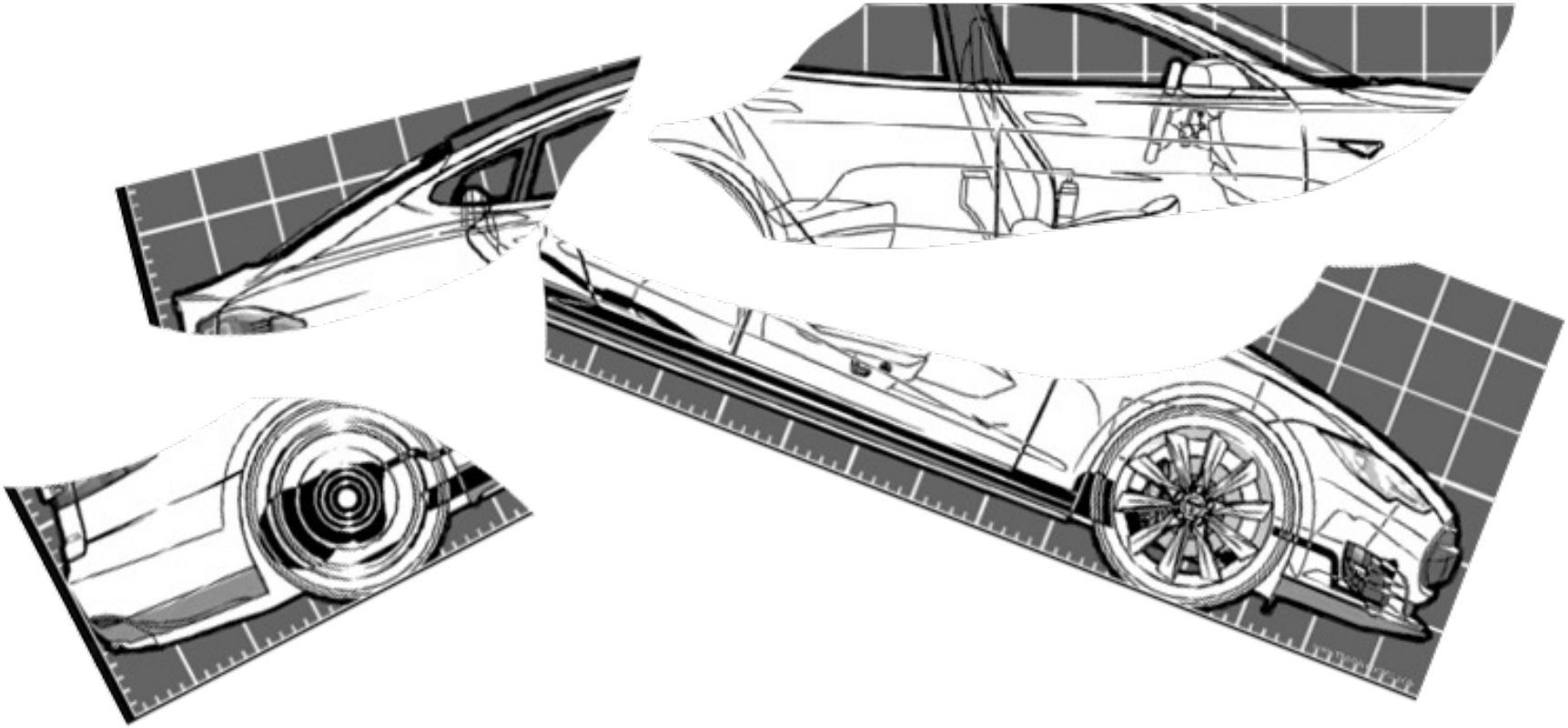
John Vollmers*, Sandra Wiegand, Florian Lenk and Anne-Kristin Kaster*

Institute for Biological Interfaces 5 (Institut für Biologische Grenzflächen IBG 5), Karlsruhe Institute of Technology (KIT) 76344, Eggenstein-Leopoldshafen, Germany

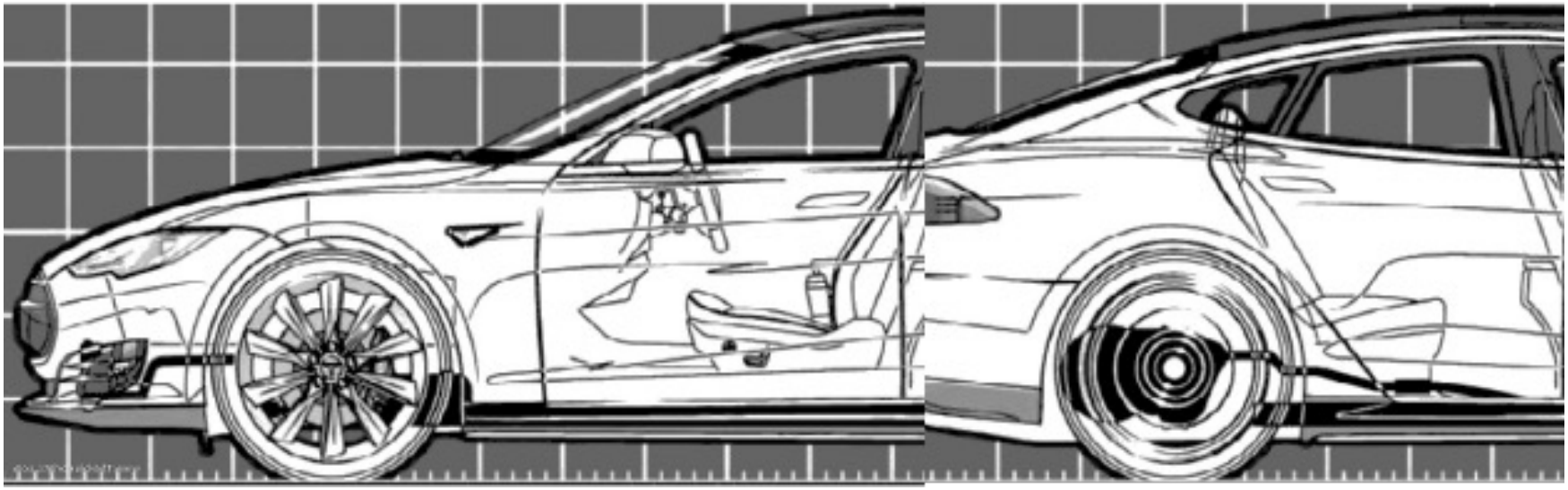
Are incompleteness and chimerism the only issues?



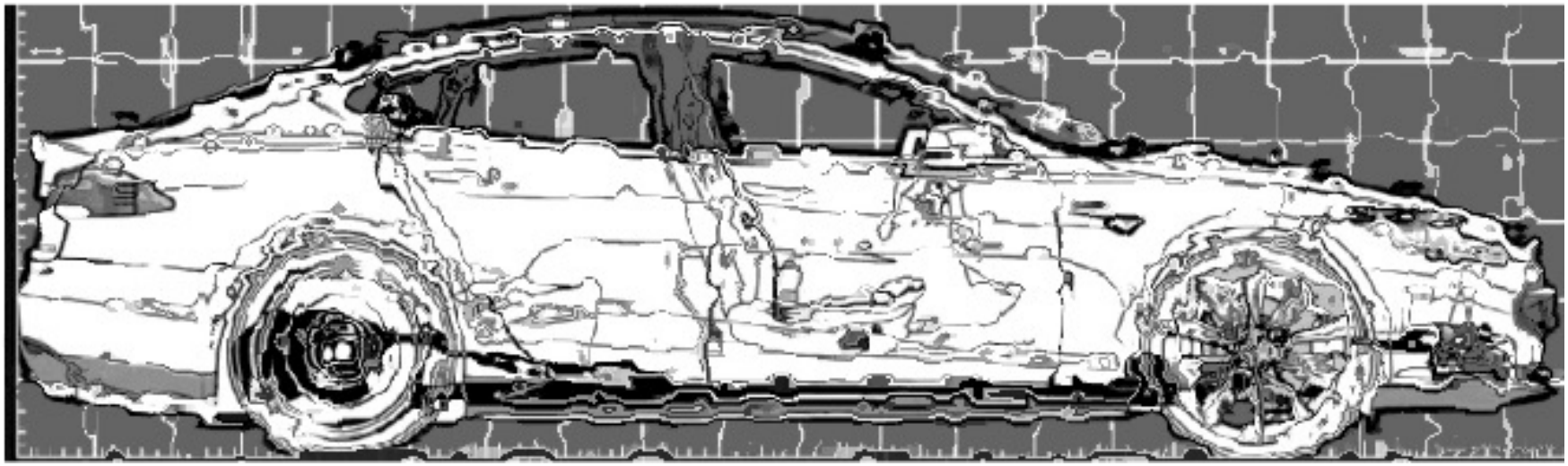
Other assembly artifacts: fragmentation



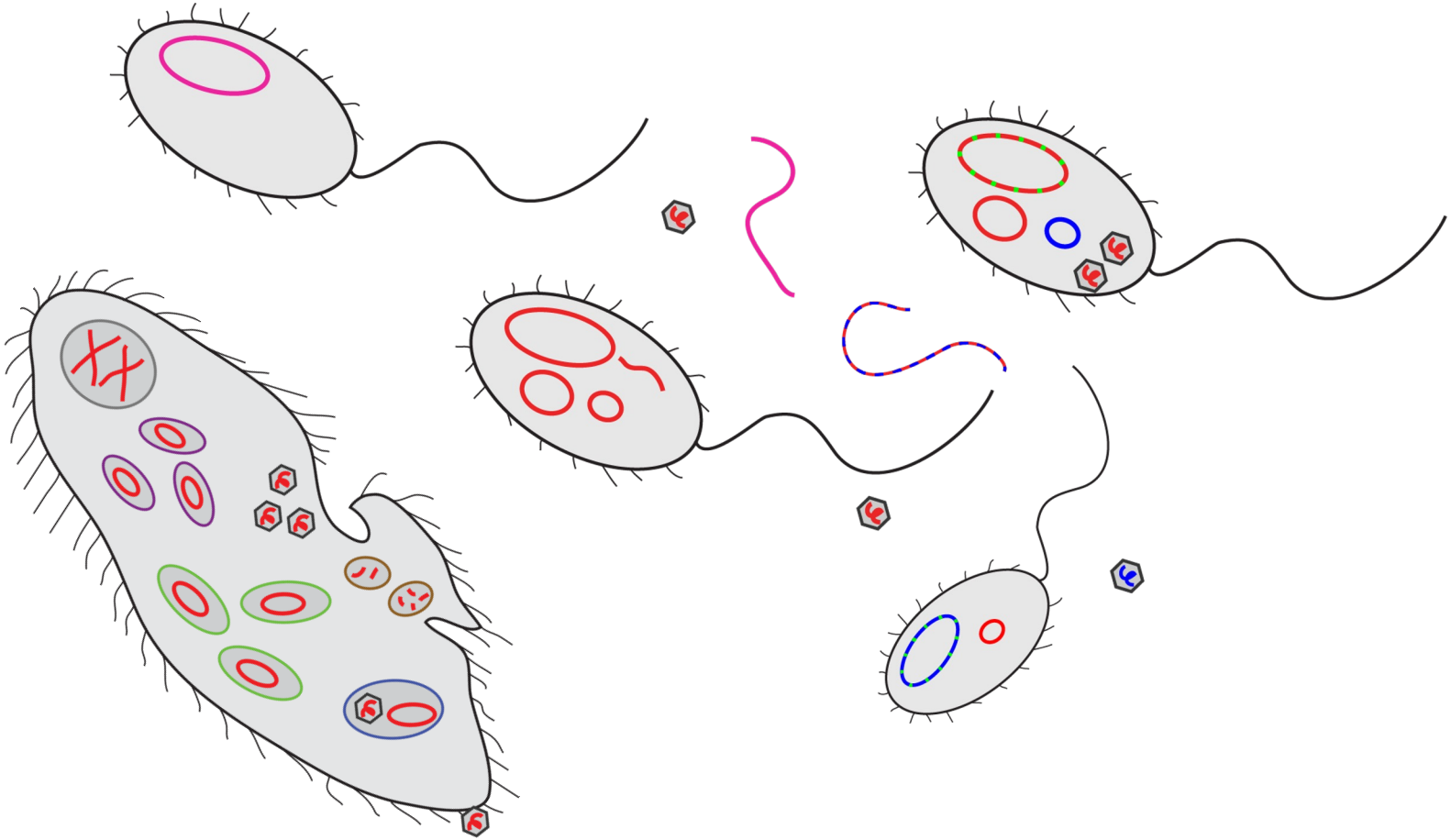
Other assembly artifacts: misassemblies



Other assembly artifacts: indels and mismatches



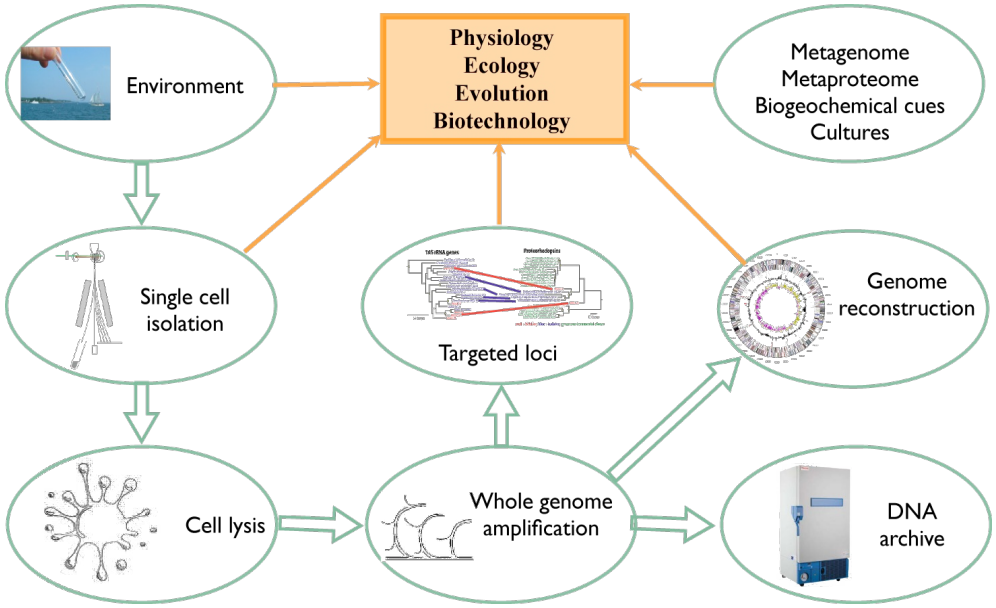
But... don’t confuse natural diversity with artifacts



QA and QC at SCGC

SCGC strategy:

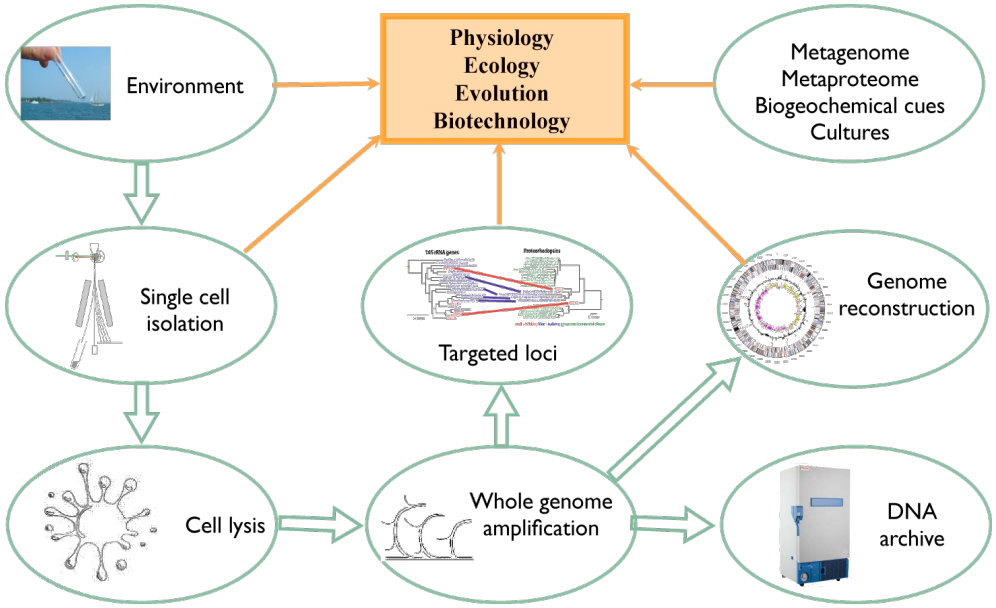
1. Validate entire workflow
2. Minimize physical DNA contamination and damage
3. Optimize accuracy of computational workflows
4. Generate data outputs for manual QC



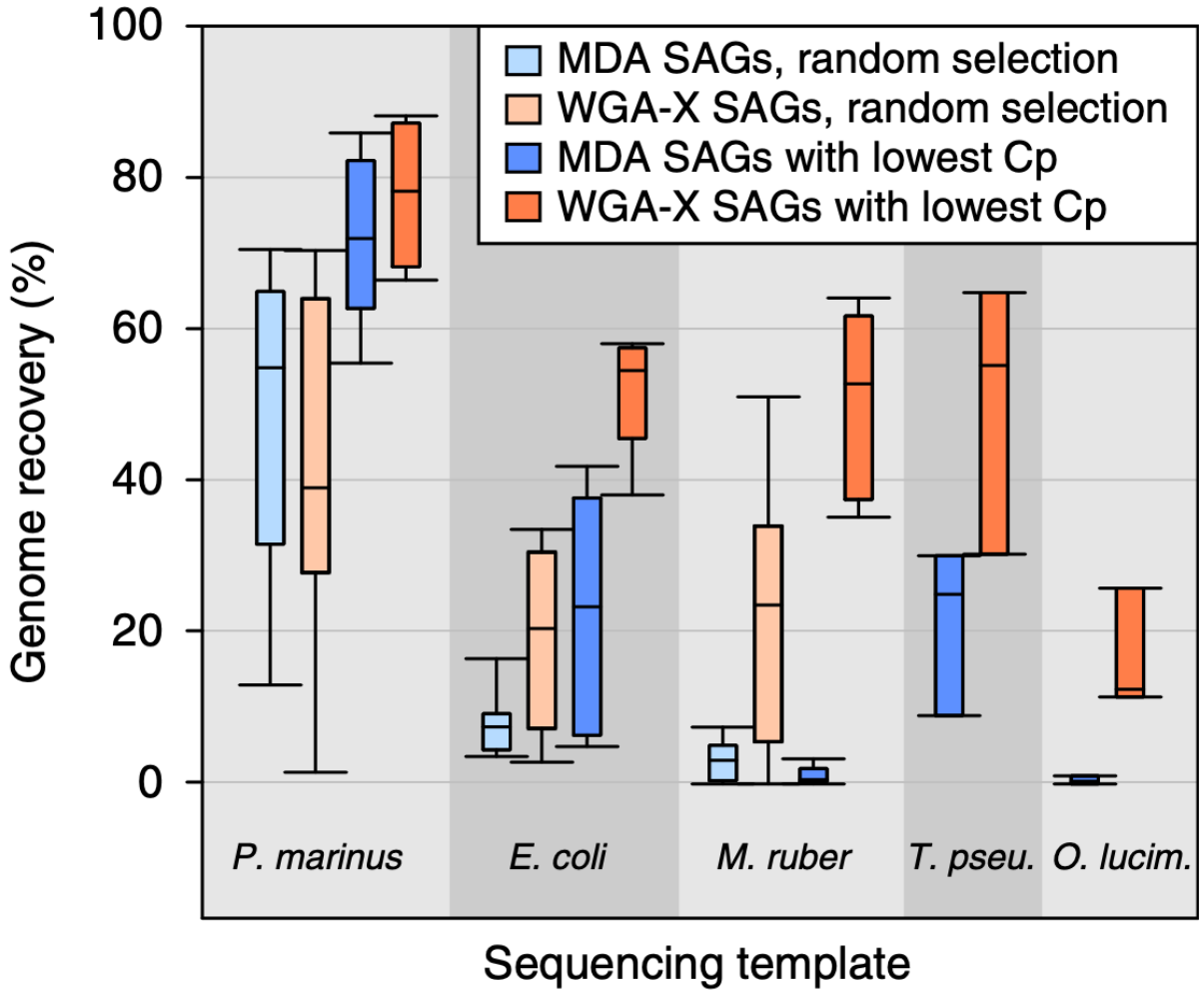
QA and QC at SCGC

SCGC strategy:

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



Validation of entire SCGC workflow



Bigelow | Laboratory for Ocean Sciences

From: Stepanauskas et al. 2017

Validation of entire SCGC workflow

BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 8 2013, pages 1072–1075
doi:10.1093/bioinformatics/btt086

Genome analysis

Advance Access publication February 19, 2013

QUAST: quality assessment tool for genome assemblies

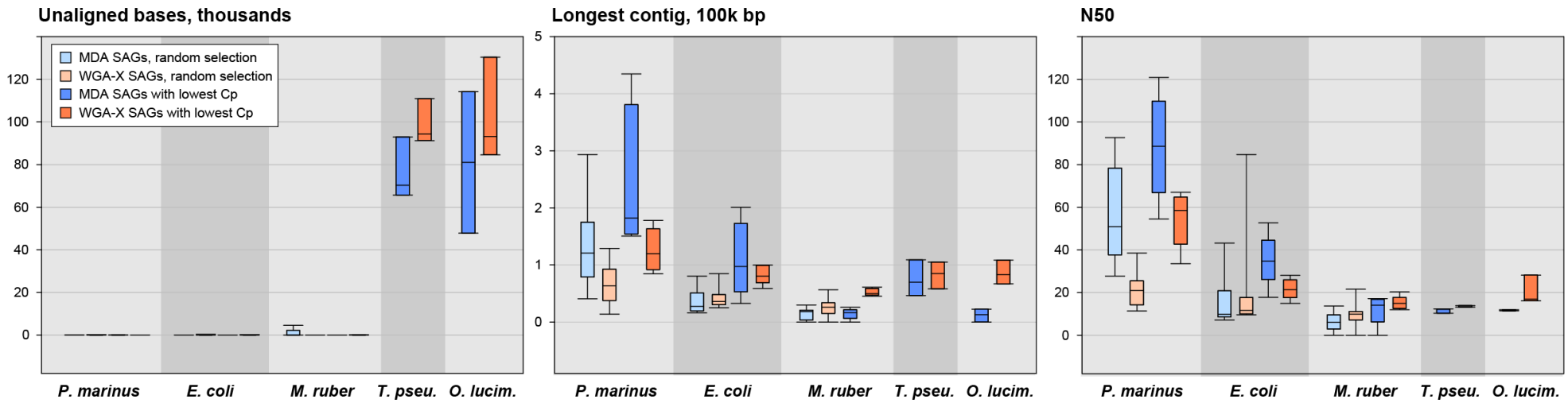
Alexey Gurevich[1,*], Vladislav Saveliev[1], Nikolay Vyahhi[1] and Glenn Tesler[2]

[1]Algorithmic Biology Laboratory, St. Petersburg Academic University, Russian Academy of Sciences, St. Petersburg 194021, Russia and [2]Department of Mathematics, University of California, San Diego, La Jolla, CA 92093-0112, USA

Associate Editor: Michael Brudno

From: Stepanauskas et al. 2017

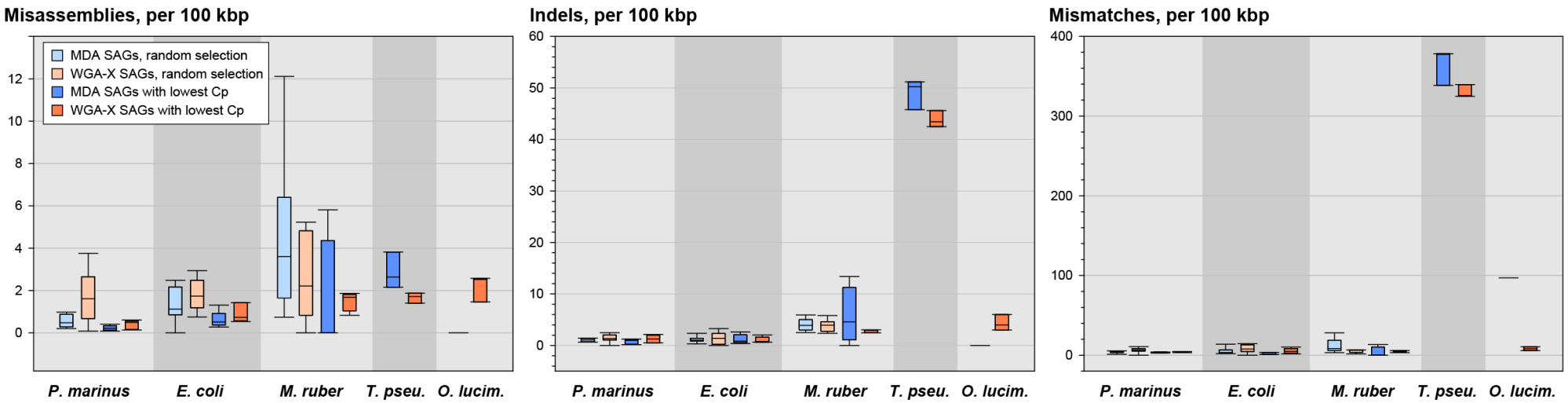
Validation of entire SCGC workflow



From: Stepanauskas et al., 2017

Bigelow | Laboratory for Ocean Sciences

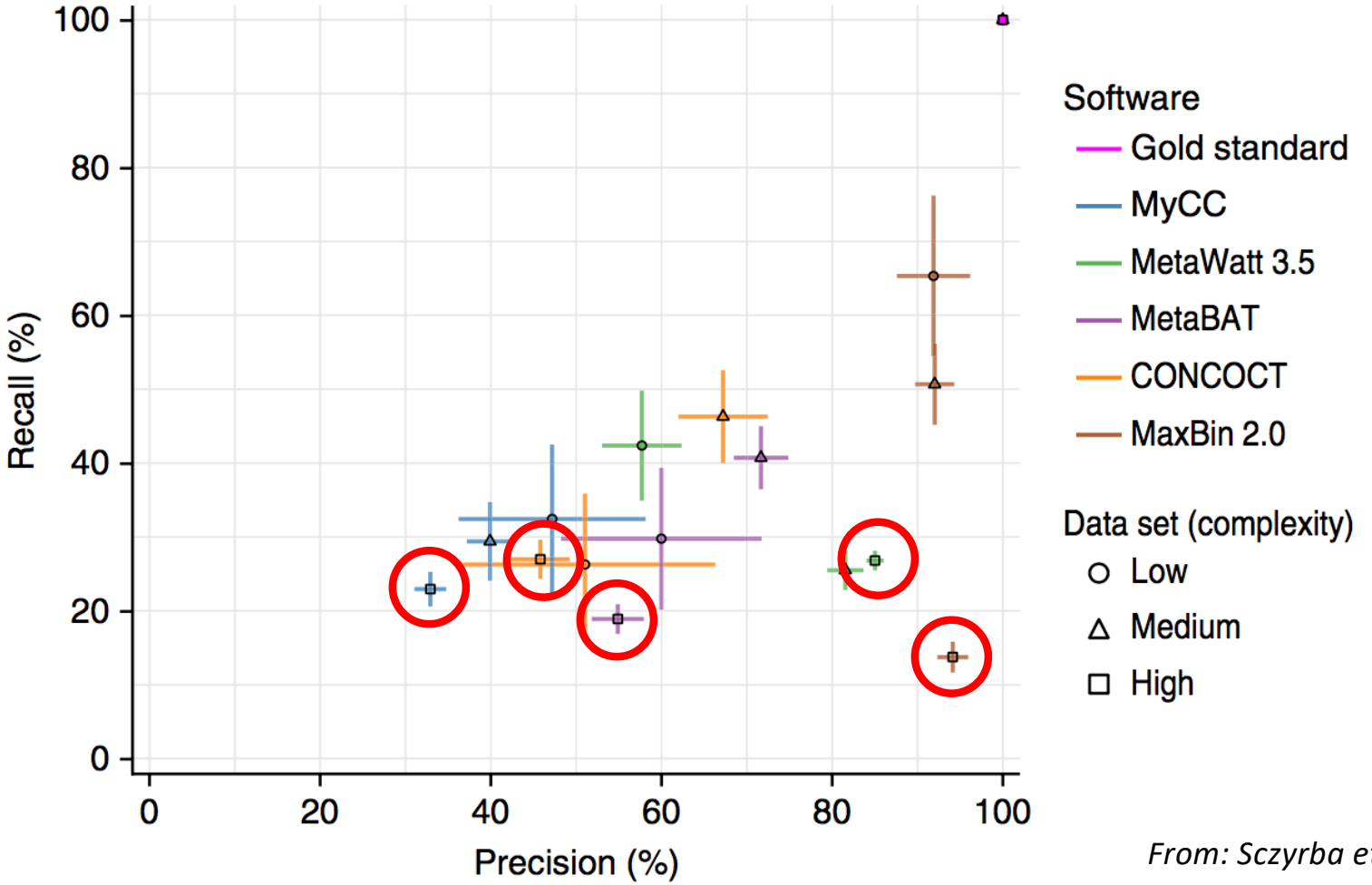
Validation of entire SCGC workflow



From: Stepanauskas et al., 2017

Bigelow | Laboratory for Ocean Sciences

Validation of MAG workflows

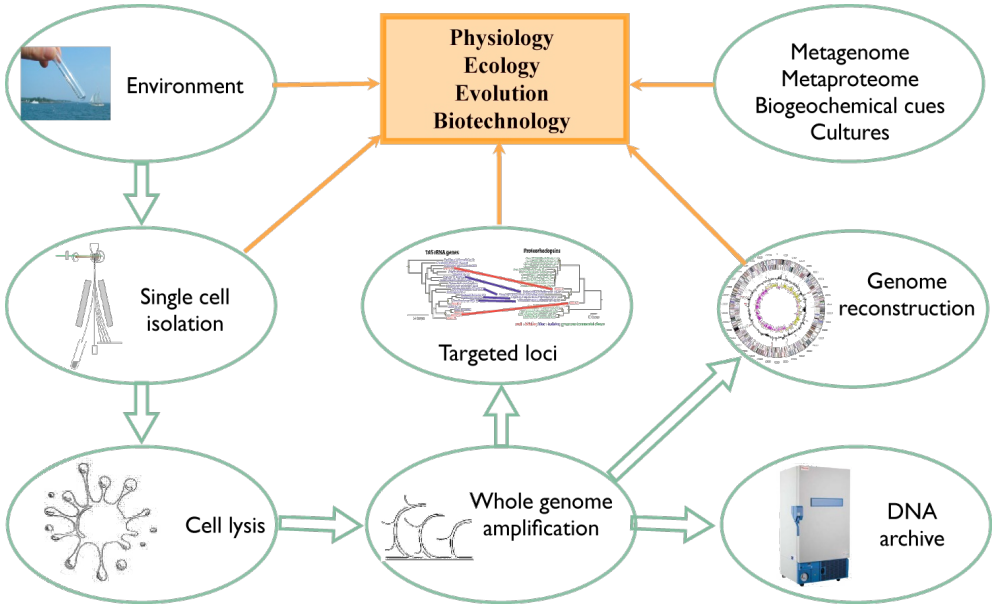


From: Sczyrba et al., Nature Methods 2017

QA and QC at SCGC

SCGC strategy:

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

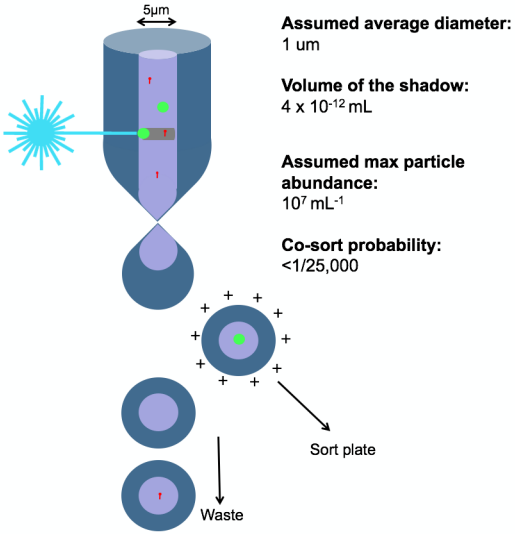


Cell sort quality assurance

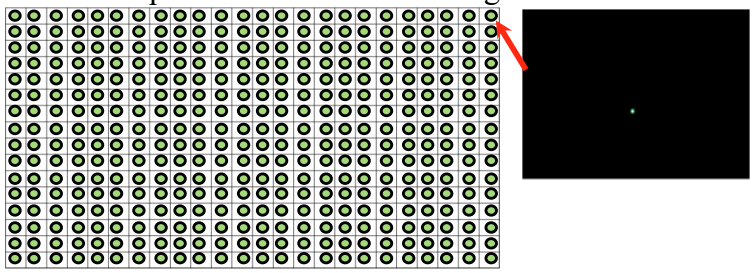
Contamination prevention:

- Prepare DNA-free sheath fluid
- Bleach sheath and sample lines
- Sort in a HEPA-filtered cleanroom
- Generate small droplets
- Include negative controls on each sort plate

Single-drop sort mode



Microscopic verification of sort alignment



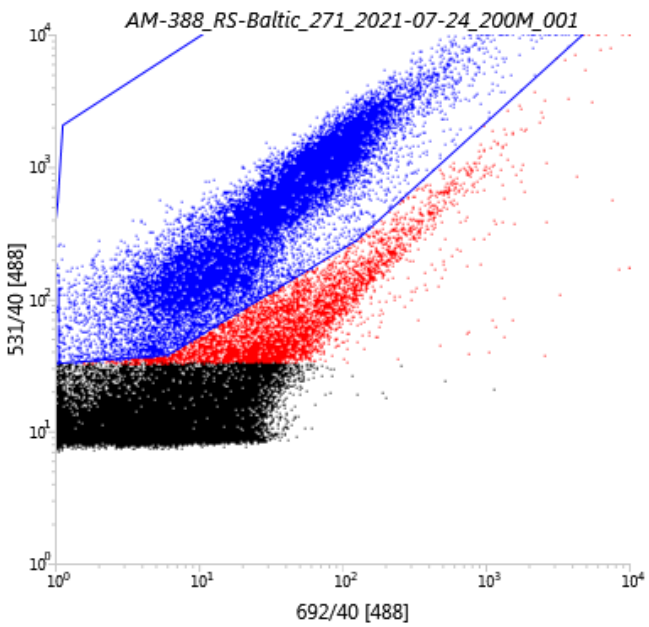
<2% wells contain no bead

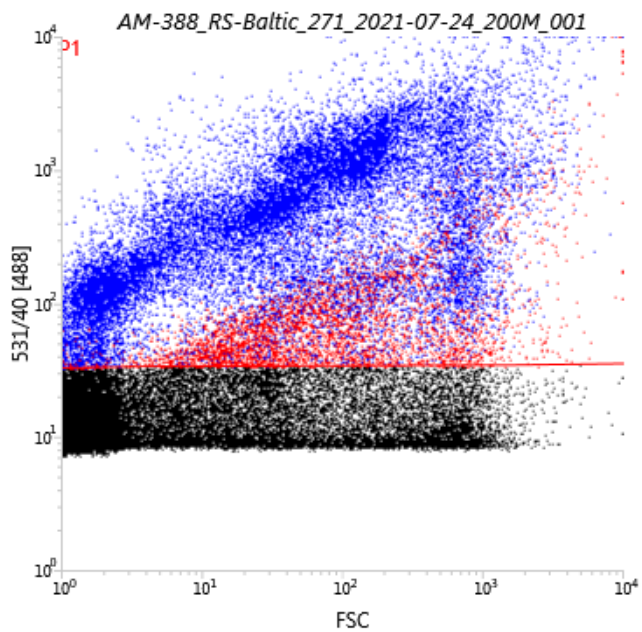
<0.4% wells contain more than one bead

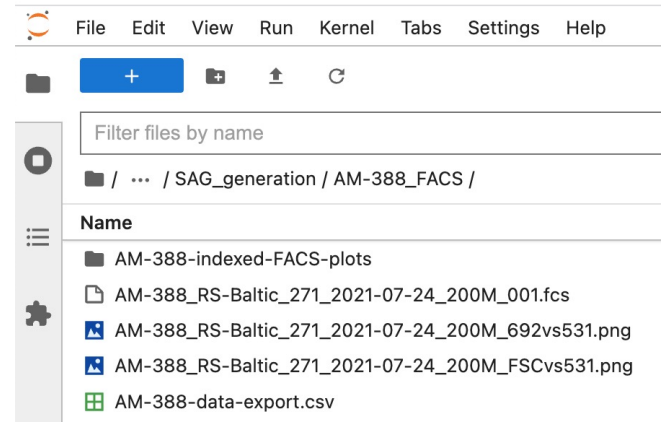
Bigelow | Laboratory for Ocean Sciences

Cell sort quality assurance

Optimize sort gate

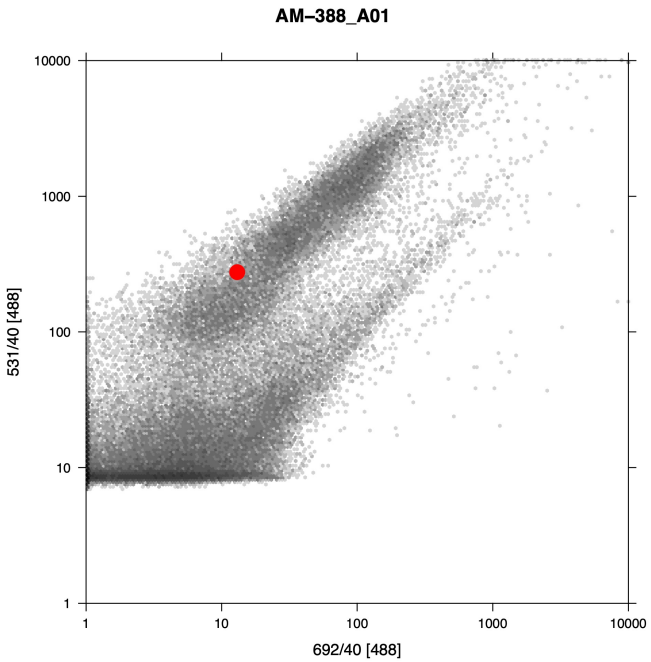


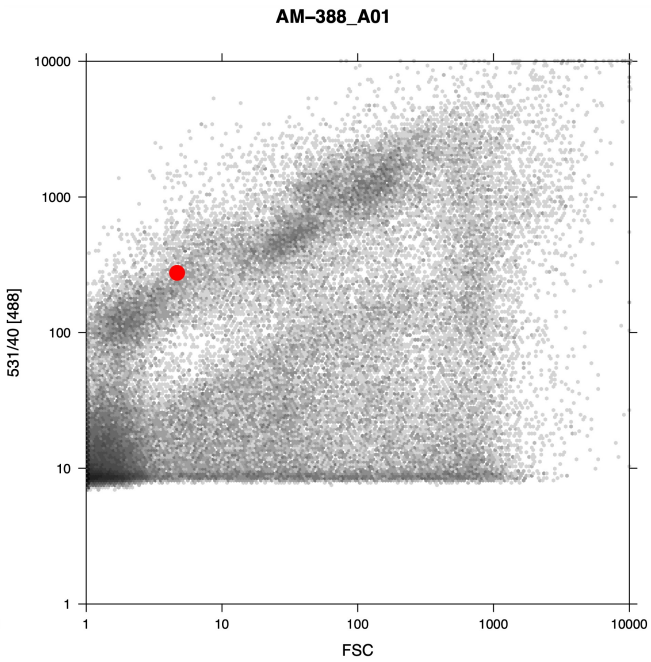


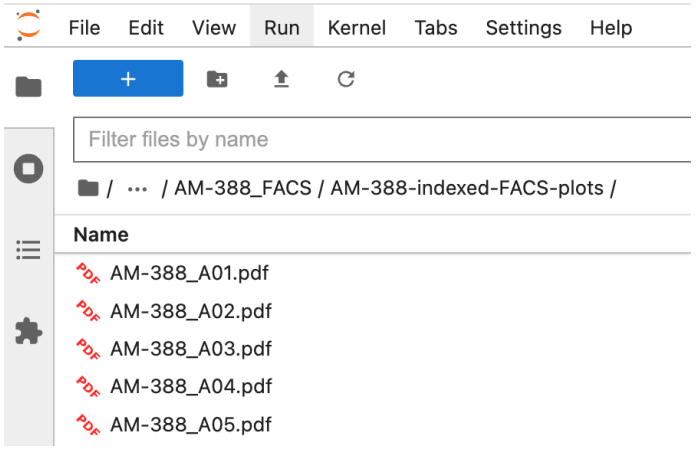


Cell sort quality assurance

Track optical properties of sorted particles







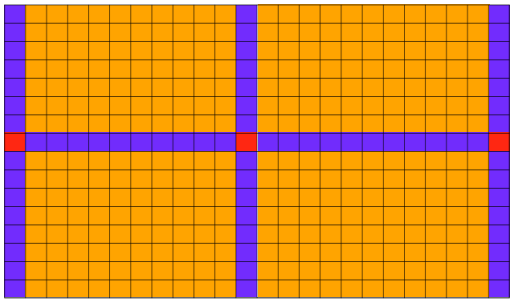
DNA amplification quality assurance

Measures for clean and efficient WGA:

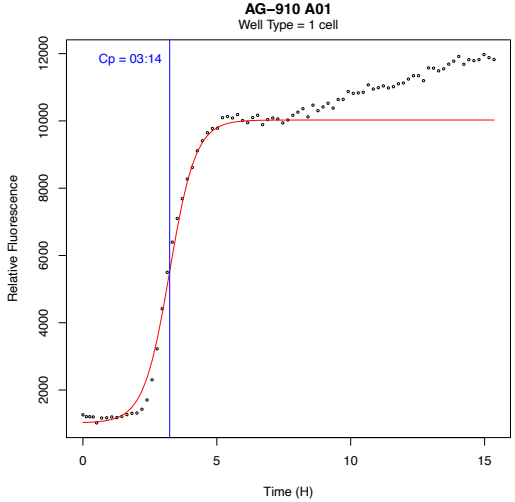
- Core method improvements: WGA-X, WGA-Y
- Robotic liquid handling in a cleanroom
- Decontamination of instruments and consumables
- Evaluation of each batch of consumables
- Negative and positive controls on each SAG plate
- Monitoring of WGA reaction kinetics

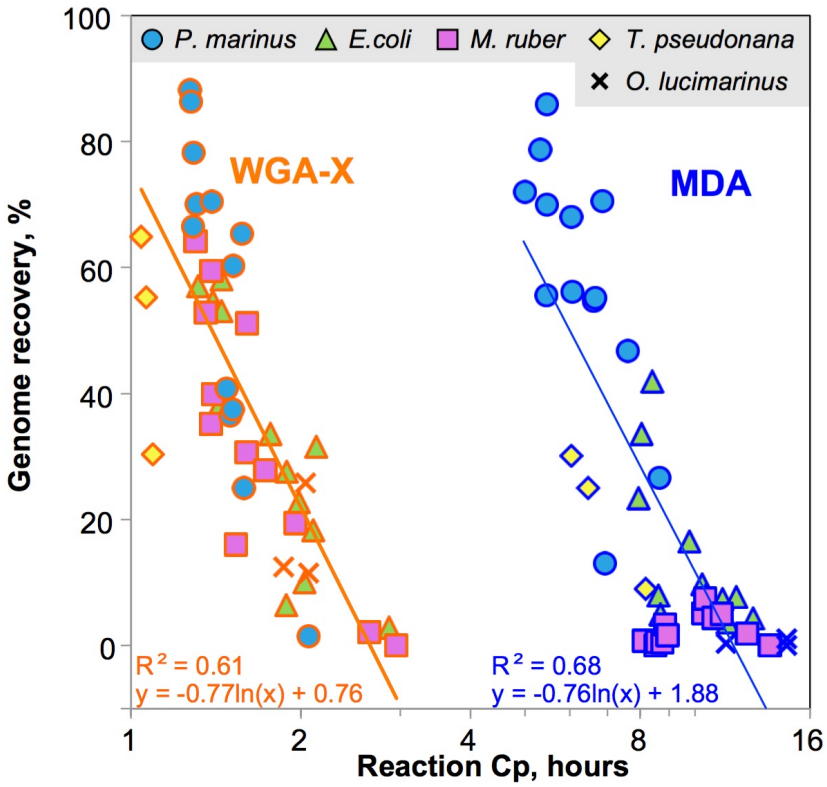
384-well plate layout

Wells containing single cells
Positive controls: 10 cells per well
Negative controls: empty wells



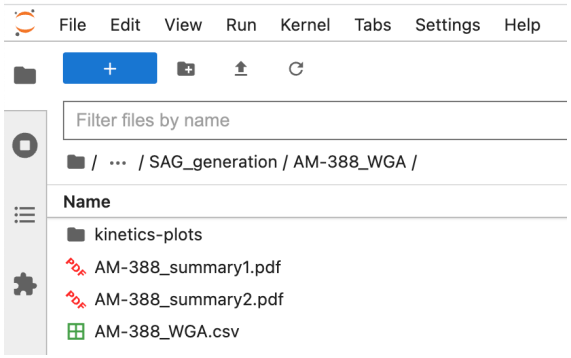
WGA Cp metric

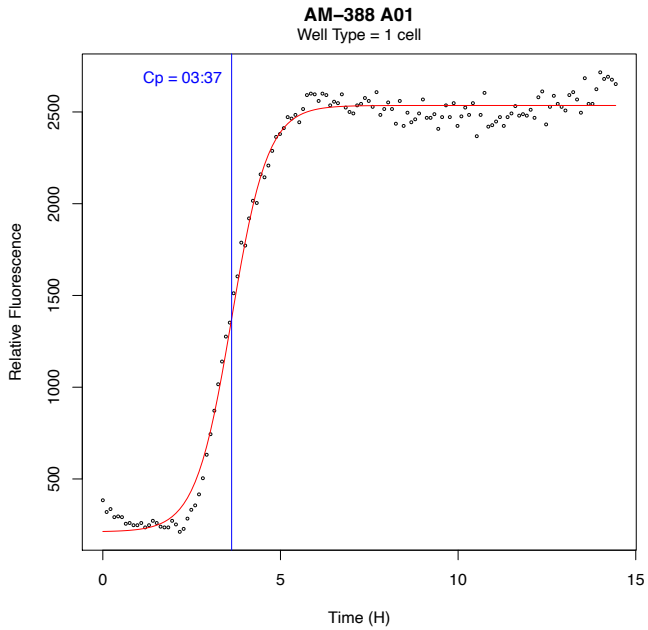


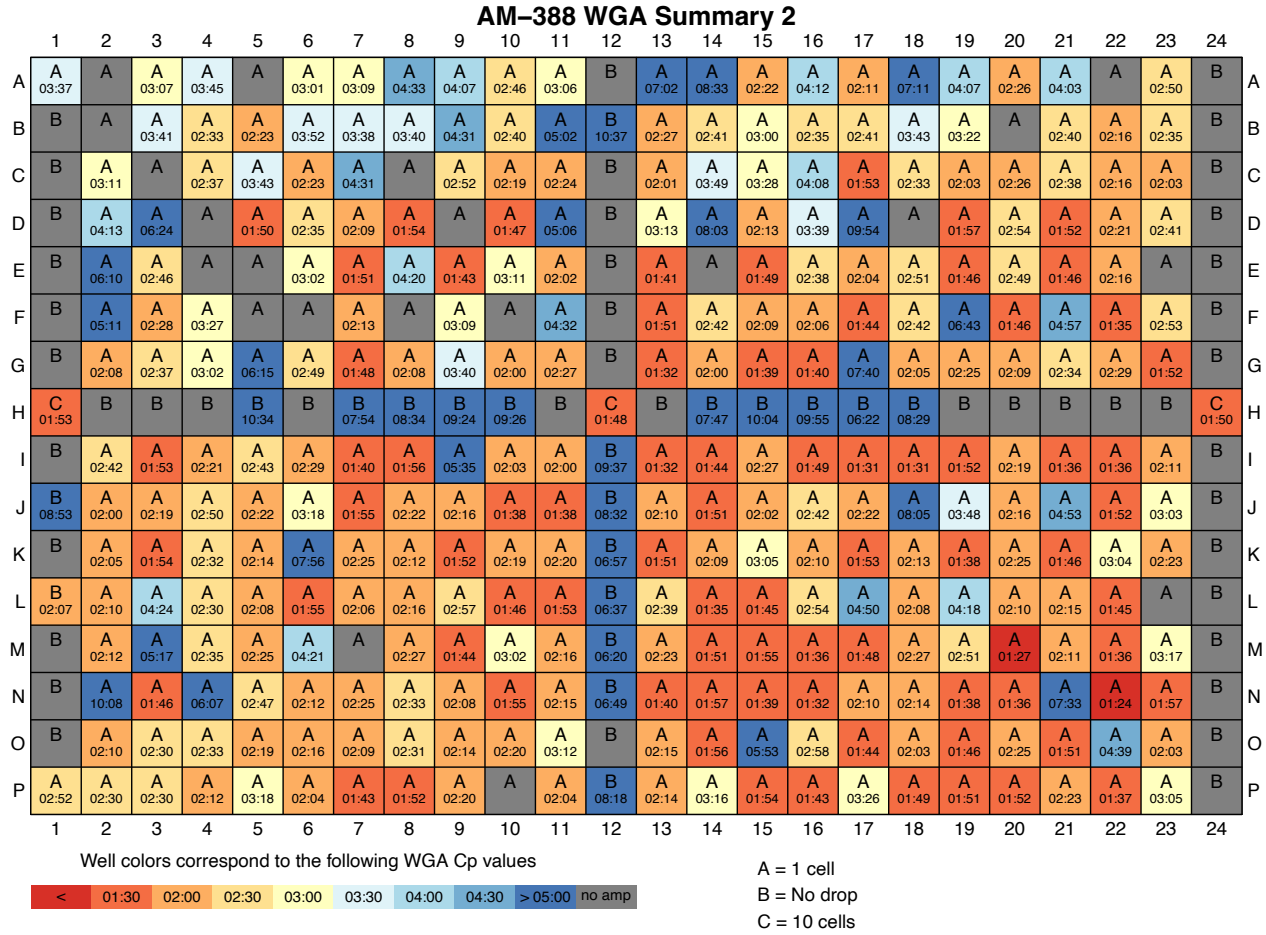


From: Stepanauskas et al. 2017

DNA amplification quality assurance



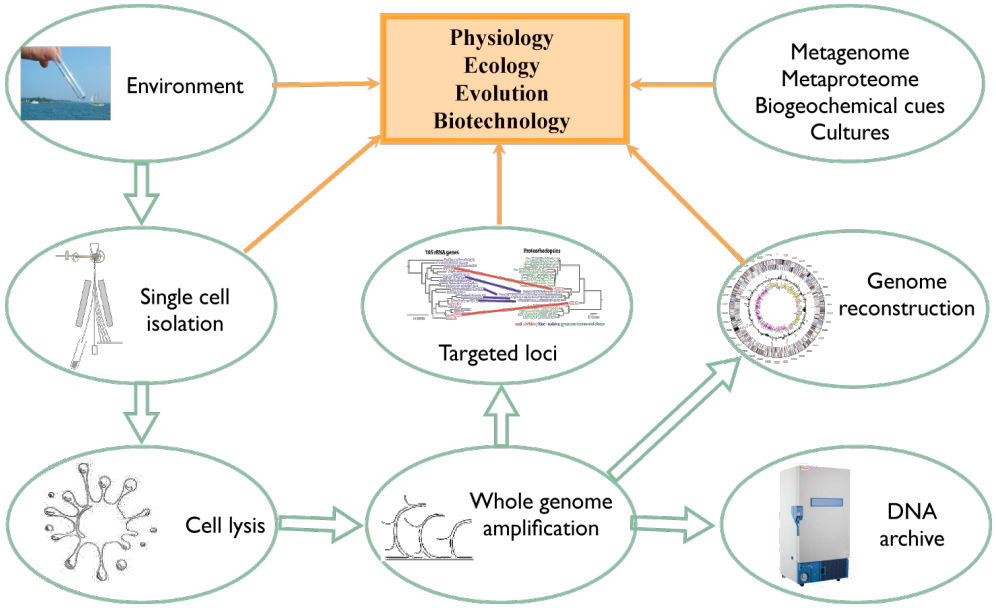




QA and QC at SCGC

SCGC strategy:

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



Optimization of computational steps

ARTICLE

DOI: 10.1038/s41467-017-00128-z OPEN

Improved genome recovery and integrated cell-size analyses of individual uncultured microbial cells and viral particles

Ramunas Stepanauskas[1], Elizabeth A. Fergusson[1], Joseph Brown[1], Nicole J. Poulton[1], Ben Tupper[1], Jessica M. Labonté[1,3], Eric D. Becraft[1], Julia M. Brown[1], Maria G. Pachiadaki[1], Tadas Povilaitis[2], Brian P. Thompson[1], Corianna J. Mascena[1], Wendy K. Bellows[1] & Arvydas Lubys[2]

Optimization of computational steps

1. Reads are quality-trimmed with Trimmomatic v0.32: -phred33 LEADING:0 TRAILING:5 SLIDINGWINDOW:4:15 MINLEN:36

1. Human DNA, reagent contaminant and low complexity reads are removed

1. Reads are digitally normalized with kmernorm using settings -k 21 -t 30 -c 3

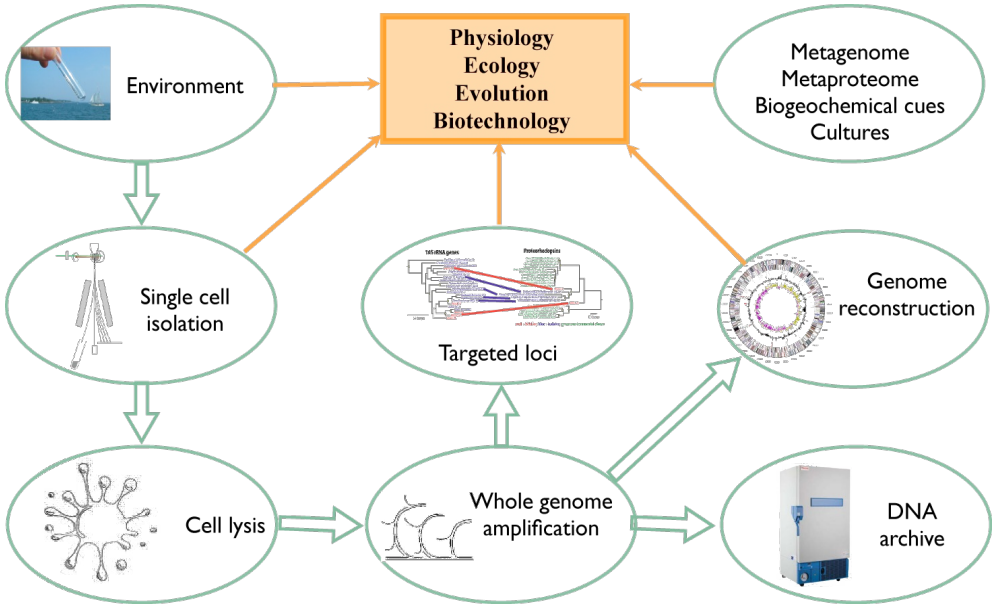
2. Normalized reads are assembled with SPAdes using the following settings: --careful --sc --phred-offset 33

3. Contigs ends are trimmed by 100 bp, and only contigs longer than 2,000 bp are retained

QA and QC at SCGC

SCGC strategy:

1. Validate entire workflow
2. Minimize physical DNA contamination and damage
3. Optimize accuracy of computational workflows
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)