Single cell genomics:
A deeper look at life’s building blocks

Ramunas Stepanauskas
Acknowledgements

SCGC staff:
Joseph Brown
Julia Brown
Wendy Bellows
Elizabeth Fergusson
Christian Harris
Monica Lluesma
Corianna Mascena
Dashiell Masland
Ben Tupper
Brian Thompson

FACS facility:
Nicole Poulton

SCGC Advisory Board
Rita Colwell
George Garrity
Kostas Konstantinidis
Craig Muir
Ger van den Engh
Tanja Woyke
Robert Young

RS postdocs (current and past):
Eric Becraft
Jane Heywood
Jessica Labonté
Manuel Garcia Martinez
Maria Pachiadaki
Brandon Swan

Key collaborators
Sallie Chisholm
Gerhad Herndl
Arvydas Lubys
Duane Moser
Stefan Sievert
Tullis Onstott
Kathuri Venkateswaran
Tanja Woyke
**Open postdoc positions**

**Scope:** Single cell genomics and horizontal gene transfer in the marine microbiome

**Application deadline:** January 15, 2017

**Further information:**
https://www.bigelow.org/about/careers.html
rstepanauskas@bigelow.org
Single cell genomics: A deeper look at life’s building blocks

Ramunas Stepanauskas

1. SCG opportunities in microbiology
2. Do we know how much we don’t know?
3. Infrastructure and method advances
Why single cell genomics?

Genomes are blueprints of life

Cells are fundamental units of biology

For most of life on Earth, 1 cell = 1 organism
Amazing features of life on Earth

- Rapid emergence
- Persistence over 1/3 of universe’s age
- Microbial dominance
Our planet may hold trillions of microbial species…

**Implications:**

- Uncultured microbes run the planet
- Only a tiny fraction will ever be cultivated
- Molecular technologies have to be pushed further

*From: Locey and Lennon 2016*
OTUs are not species!

3% divergence in the 16S rRNA gene takes ~150 million years, so...

...last common ancestor of an average OTU lived in Jurassic!

Don’t bacteria evolve faster than mammals?

Images from Wikipedia
1 mL surface ocean water encodes ~1 TB genetic information
1 g of agricultural soil encodes ~1 PB
The entire planet encodes ~10^{21} PB
<1% of this genetic resource can be accessed by cultivation
Assumptions of metagenomic assemblies

- Microbial communities are composed of clonal populations
- Entire genome evolves synchronously
- De novo assembly software does not cross-assemble different populations

Image from Wikipedia
DNA in a idealized microbial assemblage
Complex DNA distribution in a microbial assemblage

**Implications:**
Horizontal spread of antibiotic resistance and other traits
Large differences among close relatives
16S and metagenomics provide incomplete information
Microbial Single Cell Genomics Workflow
Milestones of microbial single cell genomics

Pre-2000: Experimenting with PCR-based single cell DNA amplification
- e.g. Zhang et al. (1992) PNAS 89:5847-5851

2001-2002: Development of multiple displacement amplification (MDA)
- Dean et al. (2001) Genome Research 11:1095-1099
- Dean et al. (2002) PNAS 99:5261-5266

2005-2006: Proof-of-concept single cell MDA on cultured microorganisms
- Raghunathan et al. (2005) AEM 71: 3342-3347

2007-2008: First genomic data from uncultured, single cells
- Stepanauskas and Sieracki (2007) PNAS 104: 9052-9057
- Marcy et al. (2007) PNAS 104:11889-11894

2009-now: High-throughput facilities; major research discoveries
- Over 100 publications in microbial ecology, evolution, bioprospecting and human health
Mission: make single cell genomics accessible to the broad research community; serve as an engine for discoveries in microbial ecology, evolution, bioprospecting and human health.

• First center of its kind, established 2009
• Diverse samples: aquatic, soil, organismal microbiomes, etc.
• >1,000,000 cells analyzed, representing >70 phyla
• 60+ publications since 2011, 8 in Science, Nature & PNAS
Microbiomes analyzed by SCGC
• Collected surface samples from the Gulf of Maine, the Mediterranean and the South Atlantic and North Pacific subtropical gyres in 2007-2009

• Generated >2,000 single amplified genomes (SAGs) of bacteria & archaea

• Genomically sequenced 57 SAGs representing various ubiquitous groups

• Used genomic data to analyze metabolism, biogeography and infections
Four tales by 57 marine bacterioplankton cells:

- Metabolism
- Biogeography
- Microevolution
- Interactions
Genomic divergence of cultured and uncultured bacterioplankton
%GC divergence of cultured and uncultured bacterioplankton
Four tales by 57 marine bacterioplankton cells:

- Metabolism
- Biogeography
- Microevolution
- Interactions
Global biogeography: metagenomic fragment recruitment

99% identity cutoff

% identity cutoff

60% identity cutoff
Recruitment ratio at various DNA identity intervals

ANI-16S relationship source:
Konstantinidis and Tiedje 2005, PNAS 102:2567
Divergent biogeography of close relatives
Recruitment ratio at various DNA identity intervals

16S rRNA molecular clock:

<table>
<thead>
<tr>
<th>Point</th>
<th>Event</th>
<th>Time (Myr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Diversification of Cyanobacteria</td>
<td>&gt;1300</td>
</tr>
<tr>
<td>B</td>
<td>Photosynthetic eucaryotes arise</td>
<td>&gt;800</td>
</tr>
<tr>
<td>C</td>
<td>Oxygen appears</td>
<td>&lt;2000</td>
</tr>
<tr>
<td>D</td>
<td>Oxidative eucaryotes arise</td>
<td>&gt;800</td>
</tr>
<tr>
<td>E</td>
<td>Oxygen at high concentration</td>
<td>&lt;800</td>
</tr>
<tr>
<td>F</td>
<td>Light organs appear</td>
<td>&gt;50</td>
</tr>
<tr>
<td>G</td>
<td>Eyes appear</td>
<td>&lt;500</td>
</tr>
<tr>
<td>H</td>
<td>Land plants appear</td>
<td>&lt;400</td>
</tr>
<tr>
<td>I</td>
<td>Mammals appear</td>
<td>&lt;150</td>
</tr>
<tr>
<td>J</td>
<td>Legumes appear</td>
<td>&gt;100</td>
</tr>
</tbody>
</table>

Figure from: Robert Simmon, NASA
We are still only scratching the surface
Genome content differences between two cells of SAR86 from the same drop with 100% identical 16S rRNA

The inserted region encodes a phosphate ABC transporter operon

Non-synonymous mutations
Four tales by 57 marine bacterioplankton cells:

• Metabolism
• Biogeography
• Microevolution
• Interactions
Search for viral DNA in SAGs of bacteria and archaea

- Viral marker genes
- Sequence coverage depth
- GC % and skew
- Tetramer frequency
- Recruitment of viral versus bacterial metagenomic reads
19 out of 57 SAGs (33%) contained viral sequences

10 Podoviridae phages
- Marinimicrobia SAR406
- Verrucomicrobia (5)
- Gammaproteobacteria SAR92
- Bacteroidetes
- Roseobacter

5 Myoviridae phages
- Verrucomicrobia
- Roseobacter
- Marine Group I crenarchaeon
- Marinimicrobia SAR406
- SAR86

3 Siphoviridae phages
- SAR116
- Verrucomicrobia
- Flavobacteria

1 Phycodnaviridae virus (likely contaminant)
- Verrucomicrobia

First known viruses of phyla Thaumarchaeota, Marinimicrobia, and Verrucomicrobia

Viruses were also found in SAGs of Alphaproteobacteria, Gammaproteobacteria and Bacteroidetes

Complete genome recovery of 3 phages

High fragment recruitment of viral metagenomic reads confirmed that most SAG-associated viruses are abundant in the ocean
Podoviridae phages from SAGs

DNA polymerase A phylogeny

Genomic synteny

- Synechococcus phage P60
- Synechococcus phage Syn5
- Enterobacteria phage T7
- Enterobacteria phage K1F
- Pelagibacter phage HTVC019P
- Pelagibacter phage HTVC011P
- Roseobacter phage SI01
- Celeribacter phage P12053L
- Verrucomicrobia phage AAA164-A21
- Verrucomicrobia phage AAA164-I21
- Bacteroidetes phage AAA160-P02
- SAR92 phage AAA160-D02
- Marinimicrobia phage AAA160-C11
Metagenomic fragment recruitment from viral and cellular size fractions (70% nucleotide identity cutoff)

<table>
<thead>
<tr>
<th>Bacterial metagenome (Line P)</th>
<th>Viral metagenome (POV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAG viral contigs</td>
<td>SAG host contigs</td>
</tr>
<tr>
<td>Roseobacter AAA300-J04</td>
<td>P</td>
</tr>
<tr>
<td>SAR92 AAA160-D02</td>
<td>P</td>
</tr>
<tr>
<td>Marinimicrobia AAA160-C11</td>
<td>P</td>
</tr>
<tr>
<td>Bacteroidetes AAA160-P02</td>
<td>P</td>
</tr>
<tr>
<td>Verrucomicrobia AAA164-I21</td>
<td>P</td>
</tr>
<tr>
<td>Verrucomicrobia AAA164-M04</td>
<td>P</td>
</tr>
<tr>
<td>Verrucomicrobia AAA168-E21</td>
<td>P</td>
</tr>
<tr>
<td>Verrucomicrobia AAA164-O14</td>
<td>P</td>
</tr>
<tr>
<td>Verrucomicrobia AAA164-A21</td>
<td>P</td>
</tr>
<tr>
<td>Verrucomicrobia AAA164-B23</td>
<td>P</td>
</tr>
<tr>
<td>Verrucomicrobia AAA164-P11</td>
<td>P</td>
</tr>
<tr>
<td>Roseobacter AAA076-E06</td>
<td>P</td>
</tr>
<tr>
<td>Roseobacter AAA160-J18</td>
<td>M</td>
</tr>
<tr>
<td>Marinimicrobia AAA160-I06</td>
<td>M</td>
</tr>
<tr>
<td>SAR86 AAA168-P09</td>
<td>M</td>
</tr>
<tr>
<td>Thaumarchaeota AAA160-J20</td>
<td>P</td>
</tr>
<tr>
<td>SAR116 AAA160-J14</td>
<td>P</td>
</tr>
<tr>
<td>Verrucomicrobia AAA164-A08</td>
<td>S</td>
</tr>
<tr>
<td>Verrucomicrobia AAA164-L15</td>
<td>S</td>
</tr>
<tr>
<td>Verrucomicrobia AAA164-N20</td>
<td>S</td>
</tr>
</tbody>
</table>
Possible sources of viral sequences in SAGs

1. Lysogeny
2. Lytic infection
3. Chronic infection
4. Non-infectious attachment
5. Co-sort of a cell and a free viral particle
Possible sources of viral sequences in SAGs

1. Lysogeny
2. Lytic infection
3. Chronic infection
4. Non-infectious attachment
5. Co-sort of a cell and a free viral particle
Potentially lysogenic phage in a SAR116 SAG
Possible sources of viral sequences in SAGs

1. Lysogeny

2. Lytic infection

3. Chronic infection

4. Non-infectious attachment

5. Co-sort of a cell and a free viral particle
Correlation between MDA Cp and genome recovery
Correlation between MDA Cp and genome recovery
Possible sources of viral sequences in SAGs

1. **Lysogeny**

2. **Lytic infection**

3. **Chronic infection**

4. **Non-infectious attachment**

5. **Co-sort of a cell and a free viral particle**
Possible sources of viral sequences in SAGs

1. Lysogeny
2. Lytic infection
3. Chronic infection
4. Non-infectious attachment
5. Co-sort of a cell and a free viral particle
This process is poorly understood.

Deng et al. (2012) found that non-specific attachment is rare and that most attachment is infectious.

From: Deng et al. 2012. MBio 3:e0373-00312
19 out of 57 SAGs (33%) contained viral sequences

10 *Podoviridae* phages
- Marinimicrobia SAR406
- Verrucomicrobia (5)
- Gammaproteobacteria SAR92
- Bacteroidetes
- Roseobacter

5 *Myoviridae* phages
- Verrucomicrobia
- Roseobacter
- Marine Group I crenarchaeon
- Marinimicrobia SAR406
- SAR86

3 *Siphoviridae* phages
- SAR116
- Verrucomicrobia
- Flavobacteria

First known viruses of Thaumarchaeota, Marinimicrobia, Verrucomicrobia and Gammaproteobacteria clusters SAR86 42 and SAR92

Viruses were also found in SAGs of Alphaproteobacteria and Bacteroidetes

Complete genome recovery of 3 phages

High fragment recruitment of viral metagenomic reads confirmed that most SAG-associated viruses are abundant in the ocean

1 *Phycodnaviridae* virus
- Verrucomicrobia
Possible sources of viral sequences in SAGs

1. Lysogeny

2. Lytic infection

3. Chronic infection

4. Non-infectious attachment

5. Co-sort of a cell and a free viral particle
Risk of co-sorting cells free viral particles

Make viral particles visible to the instrument

Use single-drop sort mode

Assumed cell diameter: 1 um

Volume of the shadow: 4 x 10^{-12} mL

Viral particles in 1:10 dilute seawater: 10^8 mL^{-1}

Co-sort probability: 1/2500

From Marie et al. 1999
Advantages of single cell genomics

- Independent of cultivation
- Independent of unverified microevolutionary assumptions
- Compatible with high microbial community complexity
- Embraces intracellular genetic complexity
- Requires minimal field sample quantities
- Can be integrated with single cell phenomics
Whole genome amplification kinetics

Individual reaction kinetics

Real data: *Tetraselmis* culture

Real data: marine prokaryotes from 3000 m. depth

384-well plate layout

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

16S PCR-positives
## Propensity to microbial composition biases

<table>
<thead>
<tr>
<th>Bias source</th>
<th>16S Tags</th>
<th>Shotgun metagenomics</th>
<th>Single cell genomics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filter type/ sort gate</td>
<td>Sometimes</td>
<td>Sometimes</td>
<td>Sometimes</td>
</tr>
<tr>
<td>gDNA extraction</td>
<td>High</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>gDNA amplification</td>
<td>Sometimes</td>
<td>Sometimes</td>
<td>Sometimes</td>
</tr>
<tr>
<td>Targeted PCR: primer mismatches, inserts, secondary structures, multiple gene copies</td>
<td>High</td>
<td>No</td>
<td>Low or No</td>
</tr>
<tr>
<td>DNA sequencing</td>
<td>High</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Reference databases</td>
<td>Low</td>
<td>High</td>
<td>Low</td>
</tr>
</tbody>
</table>
What do the various techniques tell about microbiomes?
Sequenced three Picobiliphyta SAGs
No evidence for autotrophy
Feed on bacteria and large viruses
Novel, picobiliphyte-infecting nanovirus
*Picomonas judraskeda* Gen. Et Sp. Nov.: The First Identified Member of the Picozoa Phylum Nov., a Widespread Group of Picoeukaryotes, Formerly Known as ‘Picobiliphytes’

Ramkumar Seenivasan¹*, Nicole Sausen¹, Linda K. Medlin², Michael Melkonian¹

¹ Department of Botany, Cologne Biocenter, University of Cologne, Cologne, Germany; ² Marine Biological Association of the UK, The Laboratory, The Citadel, Plymouth, United Kingdom
Take-home message

- We are still only scratching the surface. Current methods provide only partial and biased information about the genomic composition of microbial communities (and multicellular organisms?)

- Reduced reliance on unverified assumptions is one of the key advantages of single cell genomics, as compared to other techniques.
Mission: make single cell genomics accessible to the broad research community; serve as an engine for discoveries in microbial ecology, evolution, bioprospecting and human health.

• First center of its kind, established 2009

• Diverse samples: aquatic, soil, organismal microbiomes, etc.

• >1,000,000 cells analyzed, representing >70 phyla

• 60+ publications since 2011, 8 in Science, Nature & PNAS

Locations of SCGC customers
My lab in 2005

- Rudimentary knowledge of genomics
- No sequencing capacity
- "Not ideal" buildings
- No resources for major technology development or facility setup
What got SCGC working?

• Placing research questions first
• Team of motivated, smart people
• Tons of advice and collaborations
• Method development, continuous since 2005, ~$3 M
• Laboratory and IT equipment, ~$3 M
• Core facility business model and implementation
• QC and benchmarking of entire workflow
• Proficiency in and integration of technologies:
  • Fluorescence-activated cell sorting (FACS)
  • Whole genome amplification
  • Robotic liquid handling
  • Cleanroom environment
  • DNA sequencing
  • *De novo* assembly and other bioinformatics
  • Laboratory information management systems (LIMS)
What got SCGC working?

- Placing research questions first
- Motivated, qualified team
- A lot of advice and collaborations
- Method development, continuous since 2005, ~$3 M
- Laboratory and IT equipment, ~$3 M
- Core facility business model and implementation
- QC and benchmarking of entire workflow
- Proficiency in and integration of technologies:
  - Fluorescence-activated cell sorting (FACS)
  - Whole genome amplification
  - Robotic liquid handling
  - Cleanroom environment
  - DNA sequencing
  - De novo assembly and other bioinformatics
  - Laboratory information management systems (LIMS)
What got SCGC working?

- Placing research questions first
- Motivated, qualified team
- A lot of advice and collaborations
- Method development, continuous since 2005, ~$3 M
- Laboratory and IT equipment, ~$3 M
- Core facility business model and implementation
- QC and benchmarking of entire workflow
- Proficiency in and integration of technologies:
  - Fluorescence-activated cell sorting (FACS)
  - Whole genome amplification
  - Robotic liquid handling
  - Cleanroom environment
  - DNA sequencing
  - De novo assembly and other bioinformatics
  - Laboratory information management systems (LIMS)
SCGC LIMS – genealogy of one SAG
SCGC LIMS – FACS data
Well colors correspond to the following MDA Cp values:
SCGC LIMS – SSU rRNA PCR-sequencing data
x1,000,000+ cells
What got SCGC working?

- Placing research questions first
- Motivated, qualified team
- A lot of advice and collaborations
- Method development, continuous since 2005, ~$3 M
- Laboratory and IT equipment, ~$3 M
- Core facility business model and implementation
- QC and benchmarking of entire workflow
- Proficiency in and integration of technologies:
  - Fluorescence-activated cell sorting (FACS)
  - Whole genome amplification
  - Robotic liquid handling
  - Cleanroom environment
  - DNA sequencing
  - De novo assembly and other bioinformatics
  - Laboratory information management systems (LIMS)
Benchmarksing SAG assemblies (QUAST, 5 replicates)

**Genome characteristics**

<table>
<thead>
<tr>
<th>Organism</th>
<th>Genome size, bp</th>
<th>GC, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prochlorooccus marinus</td>
<td>1,751,080</td>
<td>37</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>4,630,707</td>
<td>51</td>
</tr>
<tr>
<td>Meiothermus ruber</td>
<td>3,097,457</td>
<td>63</td>
</tr>
</tbody>
</table>

**Standard QC metrics**

<table>
<thead>
<tr>
<th>Genome recovery, %</th>
<th>N50</th>
<th># N's per 100 kbp</th>
</tr>
</thead>
<tbody>
<tr>
<td>78</td>
<td>54,686</td>
<td>0</td>
</tr>
<tr>
<td>52</td>
<td>21,687</td>
<td>0</td>
</tr>
<tr>
<td>50</td>
<td>15,097</td>
<td>0</td>
</tr>
</tbody>
</table>

**Assessment of assembly accuracy**

<table>
<thead>
<tr>
<th>Organism</th>
<th># unaligned bases</th>
<th># misassemblies</th>
<th># local misassemblies</th>
<th># mismatches per 100 kbp</th>
<th># indels per 100 kbp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prochlorooccus marinus</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>68</td>
<td>16</td>
<td>5</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Meiothermus ruber</td>
<td>28</td>
<td>20</td>
<td>3</td>
<td>5</td>
<td>3</td>
</tr>
</tbody>
</table>

**Assumptions:**
1. Original assembly is correct.
2. No genetic changes since

**Benefits:**
1. Accurate interpretation of results
2. Continued method improvements
Factors that may impair SAG genome recovery

**Technical:**
- DNA degradation during storage/shipping
- Failure to deposit a cell into a well
- Failure to lyse the cell
- Failure to denature DNA
- Uneven WGA
- Sequencing artifacts
- Assembly artifacts
- Pipetting error at any lab step
- Computational error

**Biological:**
- DNA leaked from a mechanically damaged cell
- Cell is in a dormant state
- Host DNA was degraded by a lytic phage
- Polymerase was hijacked by small, circular plasmids
- DNA was fragmented, e.g. due to desiccation
- DNA was bound to proteins and/or other molecules, e.g. due to desiccation
- DNA was protected by intracellular compartmentalization
Contamination prevention:

- HEPA-filtered air, cleanroom techniques
- Decontamination of all reagents
- Single-drop sort mode
- Careful drop delay and sort alignment
- Negative controls on each sort plate

Small volumes is the key!

<table>
<thead>
<tr>
<th>Pressure differential</th>
<th>Ratio of $V_s$ to $V_o$ (R)</th>
<th>Sample stream diameter (µm)</th>
<th>Sample droplet volume, $V_s$ (pL)</th>
<th>ESD of $V_s$ (µm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>$3.13 \times 10^{-3}$</td>
<td>5.6</td>
<td>8.7</td>
<td>25.9</td>
</tr>
<tr>
<td>0.4</td>
<td>$4.98 \times 10^{-3}$</td>
<td>7.0</td>
<td>13.9</td>
<td>30.3</td>
</tr>
<tr>
<td>0.6</td>
<td>$7.93 \times 10^{-3}$</td>
<td>8.9</td>
<td>22.1</td>
<td>35.3</td>
</tr>
<tr>
<td>0.8</td>
<td>$9.71 \times 10^{-3}$</td>
<td>9.8</td>
<td>27.0</td>
<td>37.8</td>
</tr>
</tbody>
</table>

Microscopy of sorted fluorescent beads

<2% wells contain no bead
<0.4% wells contain more than one bead
Preventing particle co-sort

Maximize detection sensitivity

Use single-drop FACS mode

Assumed average diameter: 1 µm

Volume of the shadow: $4 \times 10^{-12}$ mL

Assumed max particle abundance: $10^7$ mL$^{-1}$

Co-sort probability: <1/25,000

Generate small droplets

<table>
<thead>
<tr>
<th>Pressure differential</th>
<th>Sample stream diameter (µm)</th>
<th>Sample droplet volume, $V_s$ (pL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>5.6</td>
<td>8.7</td>
</tr>
<tr>
<td>0.4</td>
<td>7.0</td>
<td>13.9</td>
</tr>
<tr>
<td>0.6</td>
<td>8.9</td>
<td>22.1</td>
</tr>
<tr>
<td>0.8</td>
<td>9.8</td>
<td>27.0</td>
</tr>
</tbody>
</table>

Sort plate

Waste
Whole genome amplification kinetics

**Individual reaction kinetics**

Real data: *Tetraselmis* culture

Real data: marine prokaryotes from 3000 m. depth

**384-well plate layout**

- Negative controls: empty wells
- Positive controls: 10 cells per well
- Wells containing single cells
QC of lab consumables

Clean batch of robotic pipetting tips

DNA-contaminated batch of tips
• Placing research questions first
• Motivated, qualified team
• A lot of advice and collaborations
• Method development, continuous since 2005, ~$3 M
• Laboratory and IT equipment, ~$3 M
• Core facility business model and implementation
• QC and benchmarking of entire workflow
• Proficiency in and integration of technologies:
  • Fluorescence-activated cell sorting (FACS)
  • Whole genome amplification
  • Robotic liquid handling
  • Cleanroom environment
  • DNA sequencing
  • De novo assembly and other bioinformatics
  • Laboratory information management systems (LIMS)
Excellent primers in management

- Good to Great
- Good to Great and the Social Sectors
Technology does not come first

When used right, technology becomes an *accelerator* of momentum, not a creator of it. The good-to-great companies never began their transitions with pioneering technology, for the simple reason that you cannot make good use of technology until you know which technologies are relevant. And which are those? Those—and only those—that link directly to the three intersecting circles of the Hedgehog Concept.
The Hedgehog Concept

THREE CIRCLES OF THE HEDGEHOG CONCEPT
What got SCGC working?

- Placing research questions first
- Motivated, qualified team
- A lot of advice and collaborations
- **Method development, continuous since 2005, ~$3 M**
- Laboratory and IT equipment, ~$3 M
- Core facility business model and implementation
- QC and benchmarking of entire workflow
- Proficiency in and integration of technologies:
  - Fluorescence-activated cell sorting (FACS)
  - Whole genome amplification
  - Robotic liquid handling
  - Cleanroom environment
  - DNA sequencing
  - *De novo* assembly and other bioinformatics
  - Laboratory information management systems (LIMS)
Single Cell Genomics Pipeline
**SAG cross-contamination during sequencing**

**Illumina library cross-contamination**
SAG SCGC AC-310-N17 (Firmicutes)

**Mechanisms:**
- Miss-assignment of reads among multiplexed libraries
- Sample carry-over between runs

**Solution:**
- Dual barcoding
- Use of NextSeq
- Extra care and validation of each step

**Outcome:**
- ~50 SAGs sequenced worldwide in 2013-2014
- ~1k/10k SAGs sequenced at SCGC in 2015-2016
Improved SAG de novo assembly

- **Prochlorococcus marinus CCMP1375**
  - 1,751,080 bp, %GC=36.4
  - Improvement: 3.8x

- **Escherichia coli K12 DH1 ATCC33849**
  - 4,630,707 bp, %GC=50.8
  - Improvement: 5.0x

Sequencing effort = 15 mln reads; 2x150 bp; ~4.5 Gbp
### Improved SAG de novo assembly

**Prochlorococcus marinus CCMP1375**
1,751,080 bp, %GC=36.4
*improvement: 8x*

**Escherichia coli K12 DH1 ATCC33849**
4,630,707 bp, %GC=50.8
*improvement: 13x*

Sequencing effort = 15 mln reads; 2x150 bp; ~4.5 Gbp
Improved SAG de novo assembly

**Indels per 100 kbp**

- **Prochlorococcus marinus CCMP1375**
  - 1,751,080 bp, %GC=36.4
  - Improvement: 2.0x

- **Escherichia coli K12 DH1 ATCC33849**
  - 4,630,707 bp, %GC=50.8
  - Improvement: 4.8x

Sequencing effort = 15 mln reads; 2x150 bp; ~4.5 Gbp
Improved SAG de novo assembly

Bases not aligning to reference

Prochlorococcus marinus CCMP1375
1,751,080 bp, %GC=36.4
improvement: complete elimination

Escherichia coli K12 DH1 ATCC33849
4,630,707 bp, %GC=50.8
improvement: complete elimination

Sequencing effort = 15 mln reads; 2x150 bp; ~4.5 Gbp
Improved SAG *de novo* assembly

**N's per 100 kbp**

- **Prochlorococcus marinus** CCMP1375
  - 1,751,080 bp, %GC=36.4
  - Improvement: complete elimination

- **Escherichia coli** K12 DH1 ATCC33849
  - 4,630,707 bp, %GC=50.8
  - Improvement: complete elimination

Sequencing effort = 15 mln reads; 2x150 bp; ~4.5 Gbp
Improved SAG de novo assembly

**Prochlorococcus marinus CCMP1375**
1,751,080 bp, %GC=36.4
improvement: **4.6x**

**Escherichia coli** K12 DH1 ATCC33849
4,630,707 bp, %GC=50.8
improvement: **3.8x**

Sequencing effort = 15 mln reads; 2x150 bp; ~4.5 Gbp
Improved SAG de novo assembly

**Assembly length, bp**

<table>
<thead>
<tr>
<th>Assembly length, bp</th>
</tr>
</thead>
<tbody>
<tr>
<td>3,500,000</td>
</tr>
<tr>
<td>3,000,000</td>
</tr>
<tr>
<td>2,500,000</td>
</tr>
<tr>
<td>2,000,000</td>
</tr>
<tr>
<td>1,500,000</td>
</tr>
<tr>
<td>1,000,000</td>
</tr>
<tr>
<td>500,000</td>
</tr>
<tr>
<td>0</td>
</tr>
</tbody>
</table>

- **Prochlorococcus marinus CCMP1375**
  - 1,751,080 bp, %GC=36.4
  - **loss: 22%**

- **Escherichia coli K12 DH1 ATCC33849**
  - 4,630,707 bp, %GC=50.8
  - **loss: 31%**

**Sequencing effort** = 15 mln reads; 2x150 bp; ~4.5 Gbp
Single Cell Genomics Pipeline

- Environment
  - Single cell isolation
  - Cell lysis
  - Whole genome amplification
  - DNA archive

- Physiology
  - Ecology
  - Evolution
  - Biotechnology
  - Metagenome
  - Metaproteome
  - Biogeochemical cues
  - Cultures

Targeted loci

Bigelow Laboratory for Ocean Sciences
Single Cell Genomics Pipeline

1. **Environment**
   - Single cell isolation
   - Cell lysis

2. **Physiology**
   - Ecology
   - Evolution
   - Biotechnology

3. **Metagenome**
   - Metaproteome
   - Biogeochemical cues
   - Cultures

4. **Genome reconstruction**
   - Targeted loci

5. **DNA archive**
   - Whole genome amplification
Traditional fluorescence-activated cell sorting
Gulf of Maine surface microbial community
Index cell sorting integration with single cell genomics
Gulf of Maine surface microbial community
Genomics of individual extracellular particles
Gulf of Maine surface microbial community
Commercialized systems for human single cell transcriptomics

C1 (Fluidigm)

ICELL8 (Wafergen)

Chromium (10x)  ddSEQ (BIO-RAD & Illumina)  RNA-Seq System (Dolomite Bio)
Take-home message

- Microbial single cell genomics currently requires major investment into specialized facilities

- Methods are improving rapidly for enhanced scalability, genome recovery and integration with single cell phenomics