And second

Biology is largely solved. DNA is the source code for our bodies. Now that gene sequencing is easy, we just have to read it.

It's not just "source code." There's a ton of feedback and external processing.

But even if it were, DNA is the result of the most aggressive optimization process in the universe, running in parallel at every level, in every living thing, for four billion years.

It's still just code.

OK, try opening Google.com and clicking "view source."

OK, I--... oh my god.

That's just a few years of optimization by Google devs. DNA is thousands of times longer and way, way worse.

Wow, biology is impossible.
Motivation

• Metagenomics not just for microbes – can be used to estimate insect diversity
But first...

• I would first like you to group up in groups of 3 to discuss
  – What was your favorite aspects of the Galaxy paper?

  – What weakness, if any, do you see in the Galaxy paper?

  – Brainstorm at least two additional applications of metagenomics sequencing
Great quote #1

• “Because morphological identification is precluded by the destructive nature of the collection procedure, only DNA sequence analysis is feasible making this study de facto metagenomic”
Bioinformatics methods

• Profile-based (aka binning)
  – GC content
  – K-mer content
  – Codon usage bias

• Homology-based
  – BLAST against known databases
  – Use of known markers (MetaPhyLan)
Online resources

- CAMERA
- MG-RAST
- BLAST-parsers such as MEGAN
- Online galaxy portal system
Map of the collection routes.
Steps of Galaxy pipeline

• Assess sequence quality (FastQC)

• Sequence trimming (Trimmomatic)

• Sequence comparisons (megablast; will be Metaphylan or TIPP in future)

• Phylogenetic assessment
A class-level phylogenetic profile of windshield splatter material.

Great quote #2

- The most prominent difference between the two trips is in the number of reads identified with green plants (Viridiplantae): 10,242 in trip A versus 612 in trip B. Because during each trip we collected two samples (left and right sides of the vehicle; see Methods) we were able to trace the majority (9317) of Viridiplantae reads to the left subsample. The most likely explanation for this overabundance is that a piece of plant material (e.g., a leaf or stem fragment) adhered to the collection surface.
The list included unexpected entries such as the genus Homo even though the two trips were uneventful. Such matches are likely caused by road debris (which often includes roadkill) adhering to the collecting tape.
Goals

• A complete, easy-to-use and reproducible workflow for the entire analysis

• Overcome “all vs all “ comparison on very large databases

• Collecting environmental DNA and comparing locations – very similar to YY’s project
Other reasons this is cool

• You can do it yourself. See:
  – https://usegalaxy.org/u/aun1/p/windshield-splatter

• They named their Dodge Caravan “The Wanderer”

• An excuse to visit friends!
Other applications

- National Biodefense and Analysis and Countermeasures Center (NBACC)
  - Genomics arm of Dept. of Homeland Security

- “Food” seq
Great quote

• “Total genomic DNA was extracted from 200 mg of the homogenized calibration sausages “KalD” (boiled sausage) and “KLyoA” (Lyoner sausage)
Table 3
Mapping results for the reference sausage KaD

<table>
<thead>
<tr>
<th>Species</th>
<th>Target value [%]</th>
<th>Proportion [%]</th>
<th>Difference abs. [%]</th>
<th>Difference rel. [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>AFS-quant</td>
<td>AFS-spec</td>
<td>AFS-quant</td>
</tr>
<tr>
<td>Cattle</td>
<td>35</td>
<td>36.05 ± 0.04</td>
<td>41.16 ± 0.02</td>
<td>1.05 ± 0.04</td>
</tr>
<tr>
<td>Horse</td>
<td>1</td>
<td>1.27 ± 0.01</td>
<td>1.45 ± 0.01</td>
<td>0.28 ± 0.01</td>
</tr>
<tr>
<td>Pig</td>
<td>9</td>
<td>7.22 ± 0.05</td>
<td>7.59 ± 0.09</td>
<td>1.79 ± 0.05</td>
</tr>
<tr>
<td>Sheep</td>
<td>55</td>
<td>54.76 ± 0.09</td>
<td>49.71 ± 0.08</td>
<td>0.24 ± 0.09</td>
</tr>
<tr>
<td>Waterbuffalo</td>
<td>0</td>
<td>0.64 ± 0.03</td>
<td>0.07 ± 0</td>
<td>0.64 ± 0.03</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>4 ± 0.1</td>
<td>13.38 ± 0.04</td>
<td></td>
</tr>
</tbody>
</table>

Quantitative species analysis obtained by Illumina sequencing of DNA from the "KaD" reference sausage [37]. The AFS-quant and AFS-spec approaches (see text for details) were compared. Each dataset tested contained 1 mio of paired-end sequence reads, randomly selected from a larger dataset. Three different sub-datasets (1 mio reads each) were analyzed and mean values plus standard deviations are displayed. "Difference abs." shows the difference between the proportion of reads as determined by AFS ("proportion") relative to the expected amounts existing in the sample ("target value"). "Difference rel." is calculated by dividing "Difference abs." by the expected proportion value.
Housed at -112°F in a Bern freezer is the Swiss cheesemakers’ secret weapon against forgeries. That’s where government scientists keep 10,000 strains of milk bacteria that can be used to identify copycats. The country’s cheese industry, with 604 million Swiss francs ($658 million) in exports last year, has turned to DNA fingerprinting to fight counterfeits, which Emmental producers estimate have cost them as much as 20 million Swiss francs annually.

Forgeries contribute to declining revenue for an industry already beleaguered by high production costs; and the rising Swiss franc also makes the country’s cheese more costly abroad. Exports of Emmental, for example, sank 18 percent in terms of volume in the first half of 2014. Even worse, the Switzerland Cheese Marketing industry association estimates that about 10 percent of cheese labeled as Swiss-made Emmental on supermarket shelves worldwide isn’t real.