Parallel Programming in the Classroom

Analysis of Genome Data

Karl Frinkle - Mike Morris
Parallel Programming Seminar CS4973
Spring 2015
I thought it might be appropriate to translate the title of our paper and explain what the whole exercise is about.

1. That’s why we’re here!
2. We researched the latest ACM accreditation guidelines and found that parallel computing is suggested in multiple areas of the CS curricula.
3. That is what we meant by interdisciplinary.
4. Small liberal arts colleges are notorious for not showering money on the sciences.
I thought it might be appropriate to translate the title of our paper and explain what the whole exercise is about.

1. That’s why we’re here!
2. We researched the latest ACM accreditation guidelines and found that parallel computing is suggested in multiple areas of the CS curricula.
3. That is what we meant by interdisciplinary.
4. Small liberal arts colleges are notorious for not showering money on the sciences.
This is where the 3-semester part came in.

Course 1: We needed a crash course in C that was tailored to parallel thinking, so we tweaked a standard introductory C course.
Course 2: After some MPI success it seemed like a good idea to introduce CUDA.
Course 3: After a year of fairly intense programming, we felt like students could use some validation as to why we do this in the first place. We just happened to have a real research-level math problem at hand.
Then Found LittleFe . . .
This is the Oklahoma Supercomputing Center for Research and Education’s old supercomputer, Sooner. It’s been replaced by Boomer . . .
This is the Oklahoma Supercomputing Center for Research and Education's old supercomputer, Sooner. It’s been replaced by Boomer . . .
Analyzing Genome Data
PHASE 1 – Write Code
PHASE 1 – Write Code

Definitions:

SNP: single-nucleotide polymorphism
pronounced “snip” is a DNA sequence
commonly varying within a population
PHASE 1 – Write Code

Definitions:

SNP: single-nucleotide polymorphism
pronounced “snip” is a DNA sequence
commonly varying within a population

rsid: Reference SNP cluster ID
access number used to refer
to specific SNPs
PHASE 1 – Write Code

- Harvard PGP* Database
- 23andME

<table>
<thead>
<tr>
<th># rsid</th>
<th>chromosome</th>
<th>position</th>
<th>genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs12564807</td>
<td>1</td>
<td>734462</td>
<td>AA</td>
</tr>
<tr>
<td>rs3131972</td>
<td>1</td>
<td>752721</td>
<td>GG</td>
</tr>
<tr>
<td>rs148828841</td>
<td>1</td>
<td>760998</td>
<td>CC</td>
</tr>
<tr>
<td>rs12124819</td>
<td>1</td>
<td>776546</td>
<td>AA ...</td>
</tr>
</tbody>
</table>

* PGP - Personal Genome Project
PHASE 1 – Write Code

Harvard PGP Database

Per person, there were about 1,000,000 snips.

<table>
<thead>
<tr>
<th># rsid</th>
<th>chromosome</th>
<th>position</th>
<th>genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs12564807</td>
<td>1</td>
<td>734462</td>
<td>AA</td>
</tr>
<tr>
<td>rs3131972</td>
<td>1</td>
<td>752721</td>
<td>GG</td>
</tr>
<tr>
<td>rs148828841</td>
<td>1</td>
<td>760998</td>
<td>CC</td>
</tr>
<tr>
<td>rs12124819</td>
<td>1</td>
<td>776546</td>
<td>AA</td>
</tr>
</tbody>
</table>
PHASE 1 – Write Code

- We started with 200 profiles
- Then gave ‘em names
- That was about 5G of data

Per person, there were about 1,000,000 snips.

<table>
<thead>
<tr>
<th># rsid</th>
<th>chromosome</th>
<th>position</th>
<th>genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs12564807</td>
<td>1</td>
<td>734462</td>
<td>AA</td>
</tr>
<tr>
<td>rs3131972</td>
<td>1</td>
<td>752721</td>
<td>GG</td>
</tr>
<tr>
<td>rs148828841</td>
<td>1</td>
<td>760998</td>
<td>CC</td>
</tr>
<tr>
<td>rs12124819</td>
<td>1</td>
<td>776546</td>
<td>AA...</td>
</tr>
</tbody>
</table>
!! Important Info for HPC Instructors !!

• 5G of data isn’t necessarily “big”
• Our clusters are significantly small
!! Important Info for HPC Instructors !!

- 5G of data isn’t necessarily “big”
- Our clusters are significantly small
  - We’re teaching concept and techniques
  - We can easily scale up to Boomer
PHASE 1 – Write Code

Several programs begged to be written, and all were great candidates for parallelization.

• Search for a particular rsid for a given person
• Ditto for many persons
• Both of the above for a collection of rsids
• Compare 2 persons’ makeup
  • we used a sliding window algorithm
• Compare many persons’ makeup
## Sliding Window Technique . . .

<table>
<thead>
<tr>
<th>PENEOPE KARDASHIAN</th>
<th>STONEY BURKE</th>
</tr>
</thead>
<tbody>
<tr>
<td>i6019305</td>
<td>i6019305</td>
</tr>
<tr>
<td>rs13303106</td>
<td>rs13303106</td>
</tr>
<tr>
<td>i6019306</td>
<td>i6019306</td>
</tr>
<tr>
<td>rs13303010</td>
<td>rs13303010</td>
</tr>
<tr>
<td>i6019308</td>
<td>i6019308</td>
</tr>
<tr>
<td>i6019309</td>
<td>i6019309</td>
</tr>
<tr>
<td>rs6696281</td>
<td>rs6696281</td>
</tr>
<tr>
<td>i6019310</td>
<td>i6019310</td>
</tr>
<tr>
<td>i6019311</td>
<td>i6019311</td>
</tr>
<tr>
<td>i6019312</td>
<td>i6019312</td>
</tr>
<tr>
<td>i6060381</td>
<td>i6060381</td>
</tr>
<tr>
<td>rs2340592</td>
<td>rs2340592</td>
</tr>
<tr>
<td>rs13303118</td>
<td>rs13303118</td>
</tr>
<tr>
<td>rs78164078</td>
<td>rs78164078</td>
</tr>
<tr>
<td>rs6665000</td>
<td>rs6665000</td>
</tr>
<tr>
<td>rs2341362</td>
<td>rs2341362</td>
</tr>
<tr>
<td>rs9777703</td>
<td>rs9777703</td>
</tr>
</tbody>
</table>

---

OU Supercomputer Symposium – September 23, 2015
### Sliding Window Technique...

<table>
<thead>
<tr>
<th>PENELope KARDASHIAN</th>
<th>STONEY BURKe</th>
</tr>
</thead>
<tbody>
<tr>
<td>i6019305 1 891343</td>
<td>i6019305 1 891343</td>
</tr>
<tr>
<td>rs13303106 1 891945</td>
<td>rs13303106 1 891945</td>
</tr>
<tr>
<td>i6019306 1 894379</td>
<td>i6019306 1 894379</td>
</tr>
<tr>
<td>rs13303010 1</td>
<td>894573 AA</td>
</tr>
<tr>
<td>i6019308 1</td>
<td>897792 CC</td>
</tr>
<tr>
<td>i6019309 1</td>
<td>898082 AA</td>
</tr>
<tr>
<td>rs6666281 1</td>
<td>903104 CT</td>
</tr>
<tr>
<td>i6019310 1</td>
<td>905681 CC</td>
</tr>
<tr>
<td>i6019311 1 908114 CC</td>
<td>i6019311 1 906114 CC</td>
</tr>
<tr>
<td>i6019312 1 907666 AA</td>
<td>i6019312 1 907666 AA</td>
</tr>
<tr>
<td>i6060381 1 909238 CG</td>
<td>i6060381 1 909238 CG</td>
</tr>
<tr>
<td>rs2340592 1 910935   --</td>
<td>rs2340592 1 910935 GG</td>
</tr>
<tr>
<td>rs13303118 1 918384 GT</td>
<td>rs13303118 1 918384 GT</td>
</tr>
<tr>
<td>rs78164078 1 921071 GG</td>
<td>rs78164078 1 921071 GG</td>
</tr>
<tr>
<td>rs6665000 1 924898 AC</td>
<td>rs6665000 1 924898 AC</td>
</tr>
<tr>
<td>rs2341362 1 927309 CC</td>
<td>rs2341362 1 927309 CC</td>
</tr>
<tr>
<td>rs9777703 1 928836 CT</td>
<td>rs9777703 1 928836 TT</td>
</tr>
</tbody>
</table>

If one person’s rsid was unrecorded, both were tossed out.
### Sliding Window Technique...

<table>
<thead>
<tr>
<th></th>
<th>PENELope KARDASHIAN</th>
<th>STONEY BURKE</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs13303106</td>
<td>i6019305 1 891343 GG</td>
<td>rs13303106 1 891945 AG</td>
</tr>
<tr>
<td>rs13303010</td>
<td>i6019306 1 894379 GG</td>
<td>rs13303010 1 894573 AA</td>
</tr>
<tr>
<td>i6019308</td>
<td>1 897792 CC</td>
<td>i6019308 1 897792 CC</td>
</tr>
<tr>
<td>i6019309</td>
<td>1 898082 AA</td>
<td>i6019309 1 898082 AA</td>
</tr>
<tr>
<td>rs6696281</td>
<td>1 903104 CT</td>
<td>rs6696281 1 903104 CT</td>
</tr>
<tr>
<td>i6019310</td>
<td>1 905681 CC</td>
<td>i6019310 1 905681 CC</td>
</tr>
<tr>
<td>i6019311</td>
<td>1 906114 CC</td>
<td>i6019311 1 906114 CC</td>
</tr>
<tr>
<td>i6019312</td>
<td>1 907666 AA</td>
<td>i6019312 1 907666 AA</td>
</tr>
<tr>
<td>i6060381</td>
<td>1 909238 CG</td>
<td>i6060381 1 909238 CG</td>
</tr>
<tr>
<td>rs13303118</td>
<td>1 918384 GT</td>
<td>rs13303118 1 918384 GT</td>
</tr>
<tr>
<td>rs78164078</td>
<td>1 921071 GG</td>
<td>rs78164078 1 921071 GG</td>
</tr>
<tr>
<td>rs6665000</td>
<td>1 924898 AC</td>
<td>rs6665000 1 924898 AC</td>
</tr>
<tr>
<td>rs341362</td>
<td>1 927309 CC</td>
<td>rs341362 1 927309 CC</td>
</tr>
<tr>
<td>rs9777703</td>
<td>1 928836 CT</td>
<td>rs9777703 1 928836 TT</td>
</tr>
</tbody>
</table>

---

OU Supercomputer Symposium – September 23, 2015

---

**FIRST VARIANCE**
### Sliding Window Technique . . .

<table>
<thead>
<tr>
<th>PENEOPE KARDASHIAN</th>
<th>STONEY BURKE</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs13303106 1 891945 AG</td>
<td>rs13303106 1 891945 AG</td>
</tr>
<tr>
<td>i6019306 1 894379 GG</td>
<td>i6019306 1 894379 GG</td>
</tr>
<tr>
<td>rs13303010 1 894573 AG</td>
<td>rs13303010 1 894573 AA</td>
</tr>
<tr>
<td>i6019308 1 897792 CC</td>
<td>i6019308 1 897792 CC</td>
</tr>
<tr>
<td>i6019309 1 898082 AA</td>
<td>i6019309 1 898082 AA</td>
</tr>
<tr>
<td>rs6696281 1 903104 CT</td>
<td>rs6696281 1 903104 CT</td>
</tr>
<tr>
<td>i6019310 1 905681 CC</td>
<td>i6019310 1 905681 CC</td>
</tr>
<tr>
<td>i6019311 1 906114 CC</td>
<td>i6019311 1 906114 CC</td>
</tr>
<tr>
<td>i6019312 1 907666 AA</td>
<td>i6019312 1 907666 AA</td>
</tr>
<tr>
<td>i6060381 1 909238 CG</td>
<td>i6060381 1 909238 CG</td>
</tr>
<tr>
<td>rs13303118 1 918384 GT</td>
<td>rs13303118 1 918384 GT</td>
</tr>
<tr>
<td>rs78164078 1 921071 GG</td>
<td>rs78164078 1 921071 GG</td>
</tr>
<tr>
<td>rs6665000 1 924898 AC</td>
<td>rs6665000 1 924898 AC</td>
</tr>
<tr>
<td>rs2341362 1 927309 CC</td>
<td>rs2341362 1 927309 CC</td>
</tr>
<tr>
<td>rs9777703 1 928836 CT</td>
<td>rs9777703 1 928836 TT</td>
</tr>
</tbody>
</table>

**VARIANCES**

---

OU Supercomputer Symposium – September 23, 2015
### Sliding Window Technique . . .

<table>
<thead>
<tr>
<th>PENELope KARDASHIAN</th>
<th>STONEY BURKE</th>
</tr>
</thead>
<tbody>
<tr>
<td>i6019306 1 891343 GG</td>
<td>i6019305 1 891343 GG</td>
</tr>
<tr>
<td>rs13303106 1 891945 AG</td>
<td>rs13303106 1 891945 AG</td>
</tr>
<tr>
<td>i6019306 1 894379 GG</td>
<td>i6019306 1 894379 GG</td>
</tr>
<tr>
<td><strong>rs13303010 1 894573 AG</strong></td>
<td><strong>rs13303010 1 894573 AA</strong></td>
</tr>
<tr>
<td>i6019308 1 897792 CC</td>
<td>i6019308 1 897792 CC</td>
</tr>
<tr>
<td>i6019309 1 890822 AA</td>
<td>i6019309 1 890822 AA</td>
</tr>
<tr>
<td>rs6696281 1 903104 CT</td>
<td>rs6696281 1 903104 CT</td>
</tr>
<tr>
<td>i6019310 1 905681 CC</td>
<td>i6019310 1 905681 CC</td>
</tr>
<tr>
<td>i6019311 1 906114 CC</td>
<td>i6019311 1 906114 CC</td>
</tr>
<tr>
<td>i6019312 1 907666 AA</td>
<td>i6019312 1 907666 AA</td>
</tr>
<tr>
<td>i6606381 1 909238 CG</td>
<td>i6606381 1 909238 CG</td>
</tr>
<tr>
<td><strong>rs13303118 1 918384 GT</strong></td>
<td></td>
</tr>
<tr>
<td>rs78164078 1 921071 GG</td>
<td>rs78164078 1 921071 GG</td>
</tr>
<tr>
<td>rs6655000 1 924898 AC</td>
<td>rs6655000 1 924898 AC</td>
</tr>
<tr>
<td>rs2341362 1 927309 CC</td>
<td>rs2341362 1 927309 CC</td>
</tr>
<tr>
<td><strong>rs9777703 1 928836 CT</strong></td>
<td><strong>rs9777703 1 928836 TT</strong></td>
</tr>
</tbody>
</table>

A set number of variances were allowed in a given “block” size w/o causing a “difference”.
Sliding Window Technique . . .
Sliding Window Technique . . .

This represents \textit{rsid} locations. \textit{@@} means difference. \textit{—} means agreement.
Sliding Window Technique . . .

This represents *rsid* locations. “@@” means difference. “—” means agreement.

For this example, if 2 or more do not agree in a group of 8, then record a difference. Otherwise it is recorded as a match.
Sliding Window Technique . . .

This represents \textit{rsid} locations. "@@" means difference. "—" means agreement.

For this example, if 2 or more do not agree in a group of 8, then record a difference. Otherwise it is recorded as a match.

It's a "best case" algorithm. If an \textit{rsid} is ever in a match box of 8, then it is forever a match.
Sliding Window Technique . . .

For this example, if 2 or more do not agree in a group of 8, then record a difference. Otherwise it is recorded as a match.
For this example, if 2 or more do not agree in a group of 8, then record a difference. Otherwise it is recorded as a match.

Green = Match
Red = Difference
Sliding Window Technique . . .

For this example, if 2 or more do not agree in a group of 8, then record a difference. Otherwise it is recorded as a match.

Green = Match
Red = Difference

OU Supercomputer Symposium – September 23, 2015
Sliding Window Technique . . .

For this example, if 2 or more do not agree in a group of 8, then record a difference. Otherwise it is recorded as a match.

Green = Match
Red = Difference

OU Supercomputer Symposium – September 23, 2015
For this example, if 2 or more do not agree in a group of 8, then record a difference. Otherwise it is recorded as a match.

Green = Match
Red = Difference
Sliding Window Technique . . .

For this example, if 2 or more do not agree in a group of 8, then record a difference. Otherwise it is recorded as a match.

Green = Match
Red = Difference

OU Supercomputer Symposium – September 23, 2015
Sliding Window Technique . . .

For this example, if 2 or more do not agree in a group of 8, then record a difference. Otherwise it is recorded as a match.

Green = Match
Red = Difference
Sliding Window Technique . . .

For this example, if 2 or more do not agree in a group of 8, then record a difference. Otherwise it is recorded as a match.

Green = Match
Red = Difference
Sliding Window Technique . . .

For this example, if 2 or more do not agree in a group of 8, then record a difference. Otherwise it is recorded as a match.

Green = Match
Red = Difference

OU Supercomputer Symposium – September 23, 2015
**Sliding Window Technique**

For this example, if 2 or more do not agree in a group of 8, then record a difference. Otherwise it is recorded as a match.

It’s a “best case” algorithm. If an rsid is ever in a match box of 8, then it is forever a match.

Green = Match
Red = Difference
Sliding Window Technique . . .

For this example, if 2 or more do not agree in a group of 8, then record a difference. Otherwise it is recorded as a match.

Green = Match
Red = Difference

It’s a “best case” algorithm. If an rsid is ever in a match box of 8, then it is forever a match.
Sliding Window Technique . . .

For this example, if 2 or more do not agree in a group of 8, then record a difference. Otherwise it is recorded as a match.

Green = Match
Red = Difference
Sliding Window Technique . . .

For this example, if 2 or more do not agree in a group of 8, then record a difference. Otherwise it is recorded as a match.

Green = Match
Red = Difference

OU Supercomputer Symposium – September 23, 2015
Sliding Window Technique . . .

For this example, if 2 or more do not agree in a group of 8, then record a difference. Otherwise it is recorded as a match.

Green = Match
Red = Difference
Sliding Window Technique . . .

For this example, if 2 or more do not agree in a group of 8, then record a difference. Otherwise it is recorded as a match.

Green = Match
Red = Difference

OU Supercomputer Symposium – September 23, 2015
Sliding Window Technique . . .

For this example, if 2 or more do not agree in a group of 8, then record a difference. Otherwise it is recorded as a match.

Green = Match
Red = Difference

OU Supercomputer Symposium – September 23, 2015
Sliding Window Technique . . .

For this example, if 2 or more do not agree in a group of 8, then record a difference. Otherwise it is recorded as a match.

Green = Match
Red = Difference

OU Supercomputer Symposium – September 23, 2015
Sliding Window Technique . . .

For this example, if 2 or more do not agree in a group of 8, then record a difference. Otherwise it is recorded as a match.

Green = Match
Red = Difference

OU Supercomputer Symposium – September 23, 2015
Sliding Window Technique . . .

For this example, if 2 or more do not agree in a group of 8, then record a difference. Otherwise it is recorded as a match.

Green = Match
Red = Difference

OU Supercomputer Symposium – September 23, 2015
Sliding Window Technique . . .

For this example, if 2 or more do not agree in a group of 8, then record a difference. Otherwise it is recorded as a match.

Green = Match
Red = Difference

OU Supercomputer Symposium – September 23, 2015
For this example, if 2 or more do not agree in a group of 8, then record a difference. Otherwise it is recorded as a match.
Sliding Window Technique . . .

For this example, if 2 or more do not agree in a group of 8, then record a difference. Otherwise it is recorded as a match.

Green = Match
Red = Difference

OU Supercomputer Symposium – September 23, 2015
Sliding Window Technique . . .

For this example, if 2 or more do not agree in a group of 8, then record a difference. Otherwise it is recorded as a match.

Green = Match
Red = Difference
PHASE 2 – Web GUI
<table>
<thead>
<tr>
<th>Single Person</th>
<th>Multiple People</th>
<th>Group Allele Search</th>
<th>Two Person Comparison</th>
<th>System Monitor</th>
</tr>
</thead>
</table>

**Select a person:**

LASANDRA SPPAL.

**Count No. Calls**

**Enter RSID(s) for detailed information:**

Enter RSID(s) to view detailed information.

Here are some for easy reading:

rs663200 rs1670428 rs1261555 rs528266 rs17821127 rs3657887

[Submit Query]
Currently we have 5 options.
Web GUI – Single Person

Genome Data Analyzer

Single Person | Multiple People | Group Allele Search | Two Person Comparison | System Monitor

Select a person:

IVY CONTRERES
Count no calls

Enter RSID(s) for detailed information:
(type rsid and press enter to enter more)
Here are some for easy testing: rs5986265 rs1676426 rs1253106 rs5122929 rs17641327 rs2057387

Submit Query

rs1131186 x
Web GUI – Single Person

Genome Data Analyzer

Output

CONTREJES, IVY

RSID      Allele
rs1253105  AC

No Calls
All no call variant types indicate that the sequence could not be fully resolved, either because of limited or no information, or because of contradictory information.

3828

Green is Low | Yellow is Mean | Red is High
Web GUI – Multiple Person

Genome Data Analyzer

Select some people:
- DOREEN ZAGADO
- JESUISTA SPREAVE
- CREOLA ESPINA

Enter RSID(s) for detailed information:
(RSIDs are separated by space)
rs081245 rs1676420 rs1253195 rs529529 rs17841327 rs2857807

Submit Query
## Web GUI – Multiple Person

### Genome Data Analyzer

<table>
<thead>
<tr>
<th>RSID</th>
<th>ZFACXID</th>
<th>ESRPNA</th>
<th>SPEKCEKJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs2857887</td>
<td>AA</td>
<td>AA</td>
<td>AG</td>
</tr>
</tbody>
</table>

No Calls Count: 6956, 3624, 3124

Lookup rs2857887 on snpedia.com
Web GUI – Group Allele

Enter RSID(s) for detailed information:
(This may take a few seconds as it compares all data on all nodes.
Type 'rt' and press enter to view more.
Here are some for easy testing: rs1562555 rs1357429 rs1255166 rs1229929 rs17843327 rs2857887

rs1562555 x
rs1357429 x
rs1255166 x
rs1229929 x
rs17843327 x
rs2857887 x

Submit Query
Web GUI – Two Person Compare

Select two people:
DORINE ZAGACH
MY CONRERES

Two Allele Comparison
Ignore No Calls

Window Size: 76

Max mismatches per window: 6

Mbp tolerance: 1.72

Chromosomes to compare:

Submit Query
Web GUI – Two Person Compare

Output

DOREEN ZAGACKI  IVY CONTRERES

Two Allele Comparison
5 hits matches allowed per window size of 75
only showing results with matches greater than 5 mcp in length
ignore no Calls

Selected Chromosome Output:
Web GUI – Two Person Compare

Very different – no connection. (Blue is no match.)
Web GUI – Two Person Compare

One-allele search. Blue is no match, green is match.
Web GUI – System Monitor

Cluster CPU Load

Usage:

-0.5
0.0
0.5
1.0
1.5

Time:

OU Supercomputer Symposium – September 23, 2015
Web GUI – System Monitor

Group allele – accesses all files on all nodes. Seeking 2 alleles.
Web GUI – System Monitor

Group allele – accesses all files on all nodes. Seeking 2 alleles.

Head node doing traffic only – not relatively busy.
Web GUI – System Monitor

Group allele – accesses all files on all nodes. Seeking 6 alleles.

Head node doing traffic only – not relatively busy.
Web GUI – System Monitor

Two-person compare, files on same node.
Web GUI – System Monitor

Two-person compare, files on different nodes.
Thank You!

We especially thank Henry Neeman, Charlie Peck, Tom Murphy, and everyone else associated with OU IT, the LittleFe project, the SOSU IT guys and all of our colleagues and friends in the educational community involved with HPC, for all the help we have received.

Karl Frinkle
Mike Morris

OU Supercomputer Symposium – September 23, 2015