Analysis tools and portals

Gabor Marth

Boston College

NIH Data Aggregation Meeting

Wednesday, June 5-6, 2012,
Rockville, MD

Sequences in a central place, variant recalled or aggregated, metadata and phenotype harmonized …

… what tools are needed make the data useful for the community?

What type of analyses can we do?

- Population genetic
  - Haplotype phasing
  - Single-variant allele frequencies
  - Variant burden
- Functional
  - Coding annotations
  - Disease databases
  - Non-coding annotations
  - Loss of Function analyses
- GWAS
  - Meta-analysis
  - De novo analysis
- Systems biology / higher order analyses
  - Network / pathway analysis

- Well-defined vs. open-ended
- Algorithmically easy vs. hard
- Tools mature vs. emerging
- Computation-heavy vs. light

Who are we trying to serve?

How does my tool perform?

Statistician / tool developer

What's different about the non-responder?

Drug developer

Is this variant associated with any known phenotype?

Biologist in small laboratory

Do I get a bigger p-value from all the extra samples?

Medical consortium project analyst

I see a variant in this gene… should I alter the treatment of my patient?

Clinician

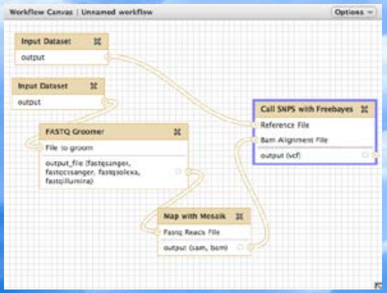
How to make analysis accessible?

- Easy to install
- Easy to use
- Intuitive
- Fast
- Interactive
- Web-based

- Storage & hardware
- Informatics expertise

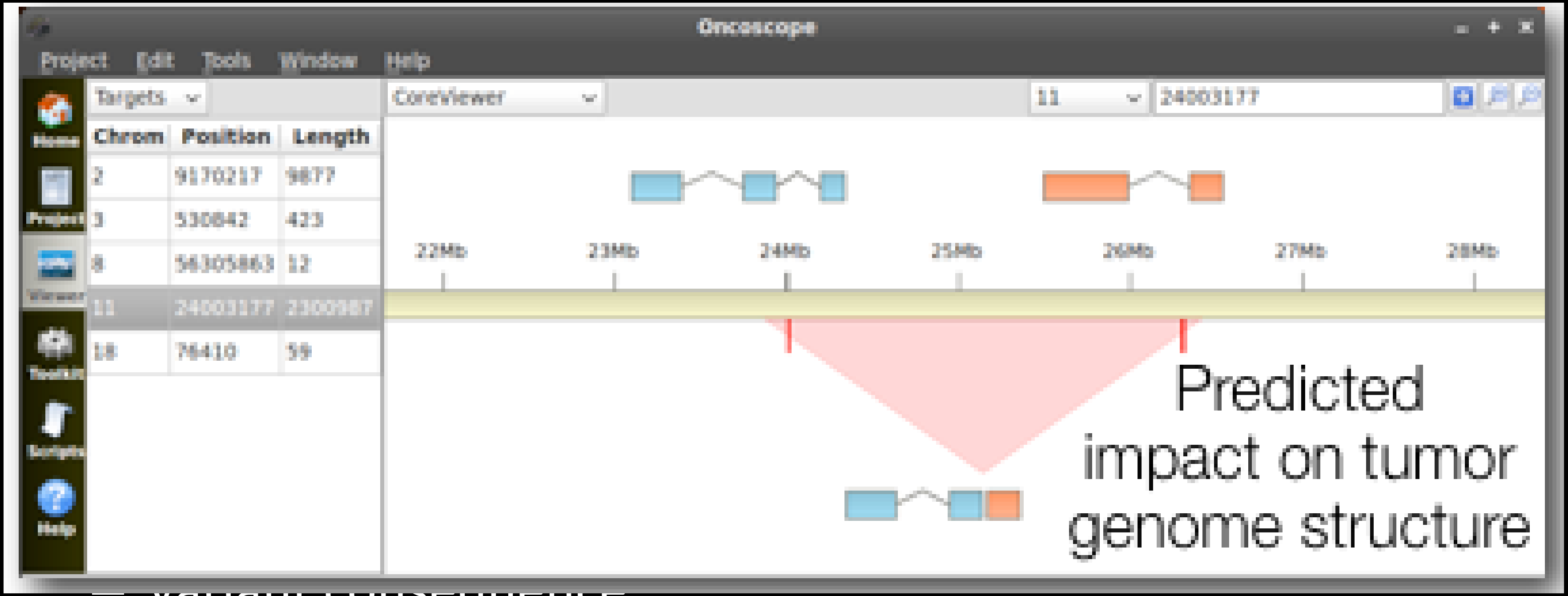
How can we provide the analyses?

- Raw data download
- Query portals, viewers, data slicers
- Static variant annotations, pre-computed resources
- Analysis environment with **central data**, facilities for users to add their **own data**, **tools installed** and **computational resources** to run the analysis



What static analyses make sense?

- Variants, variant allele frequencies, sample genotypes
- Phased haplotypes



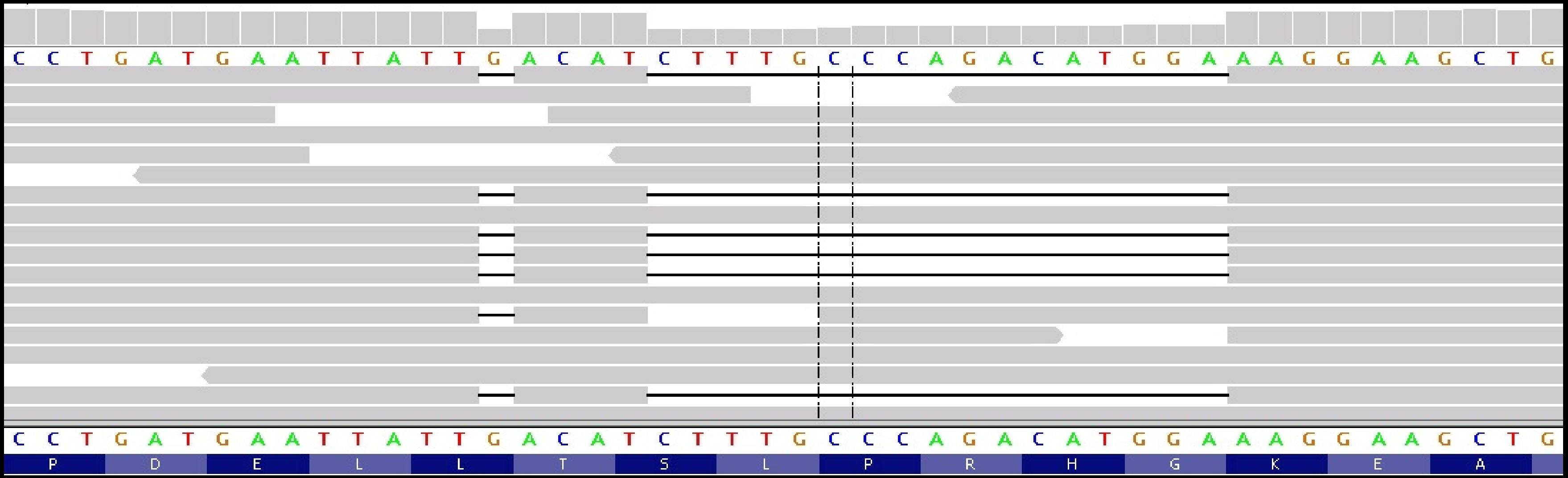
Chrom	Position	Length
2	9170217	9877
3	530842	423
8	56305863	12
11	24003177	2300987
18	76410	59

Predicted impact on tumor genome structure

  - Variant consequence
  - Haplotype browsing
  - Phenotype browsing

What tasks require analysis services?

- Read mapping, and variant calling


C C T G A T G A A T T A T T G A C A T C T T T G C C C A G A C A T G G A A A G G A A G C T G
C C T G A T G A A T T A T T G A C A T C T T T G C C C A G A C A T G G A A A G G A A G C T G
P D E L L T S L P R H G K E A

(Daniel McArthur)

- This category has the highest tool development cost because of the additional engineering required

One tool or multiple tools?

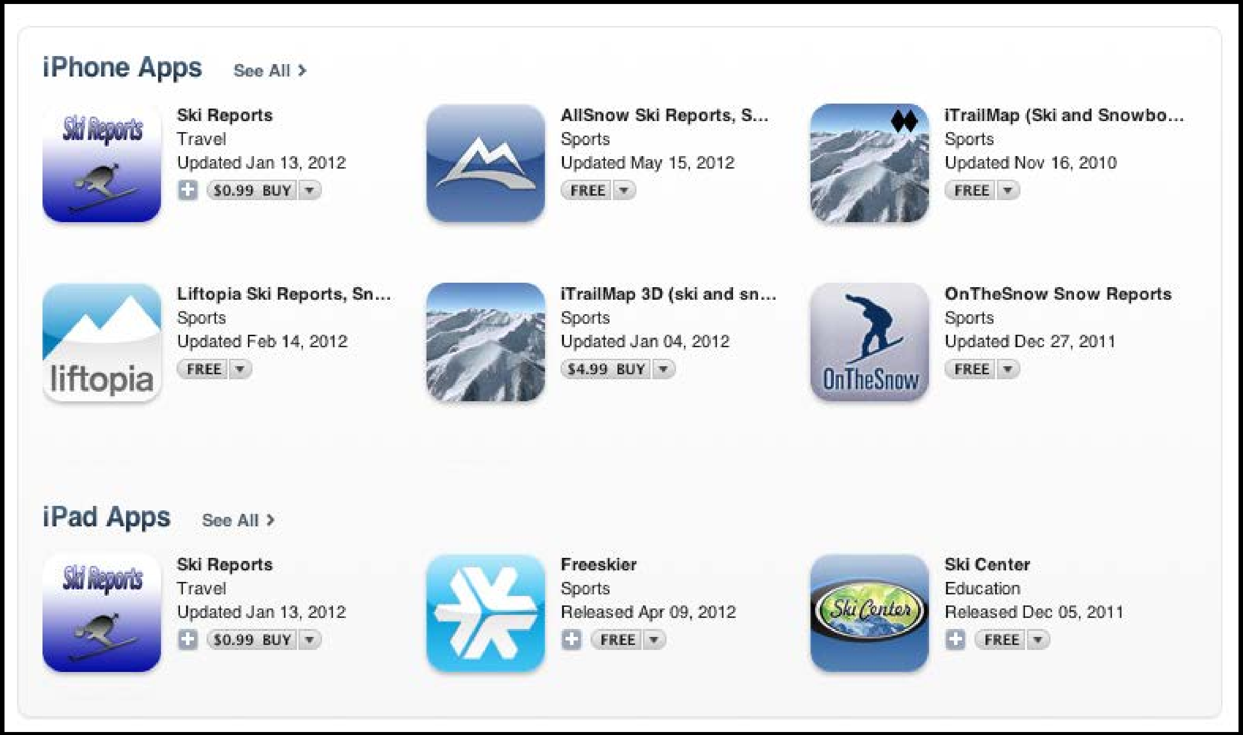
		FP metric
Dataset	SNPs	Total
Tool 1	632,344	2.32
Tool 2	547,173	2.34
Tool 3	576,125	2.36
Tool 4	629,761	2.26

		FP metric
Dataset	SNPs	Ts/Tv
4 of 4	410,243	**2.56**
3 of 4	518,407	**2.50**
2 of 4	593,538	**2.42**

There are inherent advantages to having alternative tools available

Centralized or distributed development?

- Tool development is iterative… once we get an answer, we want to ask a new question
- Often users are better served by light, flexible tools for customized analysis… a **tool ecosystem**



iPhone Apps
See All >

Ski Reports
Travel
Updated Jan 13, 2012
$0.99 BUY

AllSnow Ski Reports, S...
Sports
Updated May 15, 2012
FREE

iTrailMap (Ski and Snowbo...
Sports
Updated Nov 16, 2010
FREE

Liftopia Ski Reports, Sn...
Sports
Updated Feb 14, 2012
FREE

iTrailMap 3D (ski and sn...
Sports
Updated Jan 04, 2012
$4.99 BUY

OnTheSnow Snow Reports
Sports
Updated Dec 27, 2011
FREE

iPad Apps
See All >

Ski Reports
Travel
Updated Jan 13, 2012
$0.99 BUY

Freeskier
Sports
Released Apr 09, 2012
FREE

Ski Center
Education
Released Dec 05, 2011
FREE

Who would develop the tools?

- Many current tools from large genome centers but the majority from smaller tool development groups

- There is also a large and successful “cottage industry” of tool development, where small informatics groups can produce very sophisticated software, and respond nimbly to user needs

How to move forward?

- Focus on the cloud
- Build an open environment for tool deployment to pull in the widest possible developer base
- Models and technologies exist (iPhone apps, etc.)