Preliminary Discussion

SKCM as a TestBed for TCGA End-To-End Reproducibility

M. Noble
Broad Institute GDAC
TCGA SKCM Telecon
May 14, 2013

Review: May 2013 Steering Committee in Seattle

Approved Policy:

Scripts used to generate levels 2, 3, and 4 data should be made available with the data, as should version information for both scripts and data. At minimum, the algorithms should be documented in sufficient detail that they can be implemented by external users.

Thus Spoketh Kenna

Review: cont.

"We can use Melanoma ..."

Thus Respondeth Lynda

So, what exactly does this all mean?

Approved Policy:

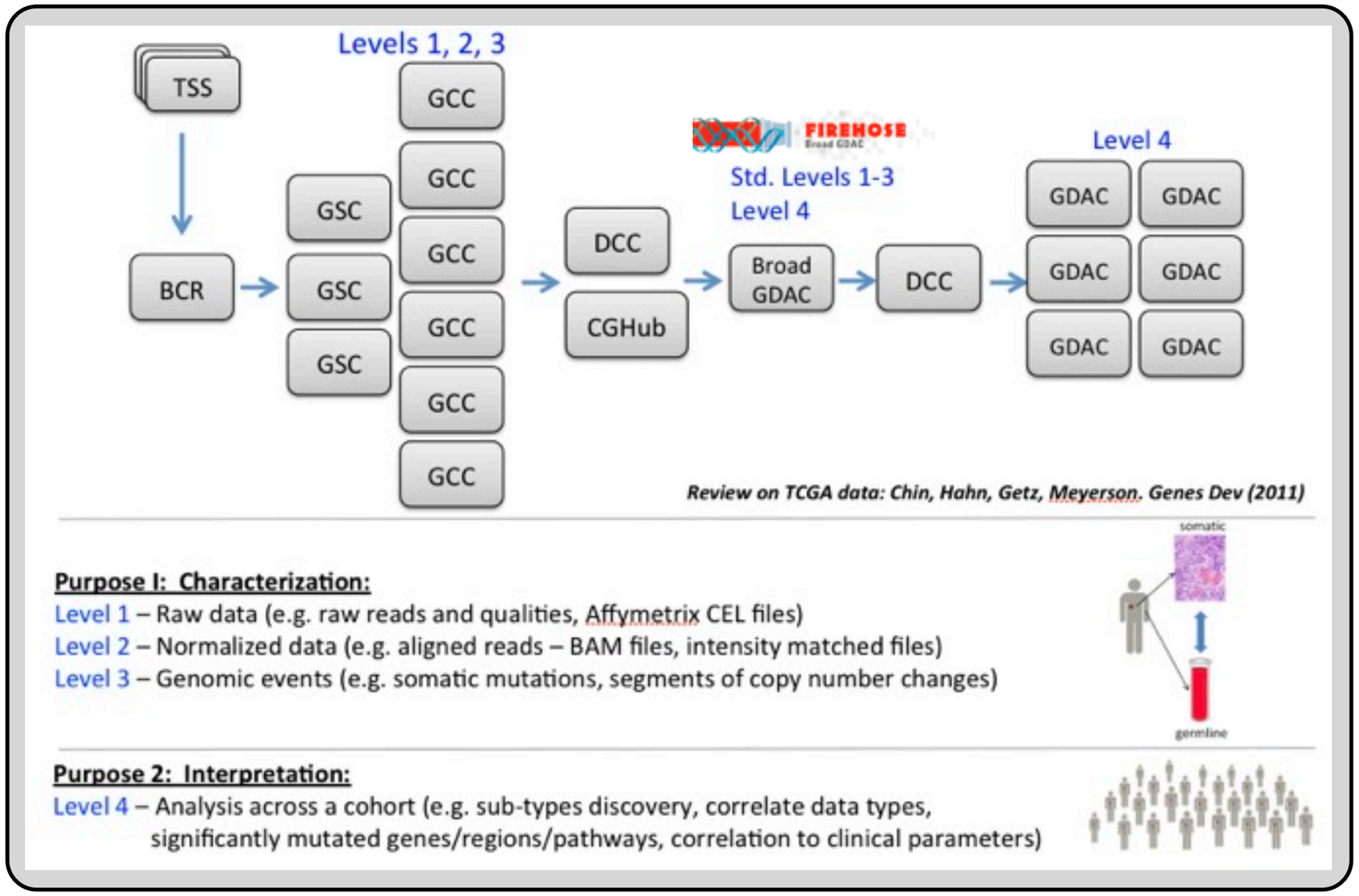
Scripts used to generate levels 2, 3, and 4 data should be made available with the data, as should version information for both scripts and data. At minimum, the algorithms should be documented in sufficient detail that they can be implemented by external users.

And clinical (level 1)

Recommendation 1: clarify scope

Recommendation 2: define clear metric of success

Refresh: From Tissue to Publication



AWG

nature

This shows GDAC Firehose a good start:

Centralized nexus of most analyzable data
Largely automatable:

linux% fiss sample_list awg_skcm__2013_MM_YY

But far from complete solution:
Not yet open for arbitrary use
Data & analysis holes remain to be filled

Hole: BAMs? (stored outside of DCC)

We need trace-ability back to source BAMS

EG: this came up in PANCAN group

We're developing script @ Broad

To add source BAM as extra column in freeze table

Other characterization data ok
(stored at DCC, mirrored in Firehose)

BUT ....

Hole: sequencing & characterization algorithms

Picard? Aligners?
Mutect (SKCM sequenced at Broad)
SNP6 pipeline
Methylation
miR-Seq mRNA-Seq
RPPA
Low-Pass CN?

Meta versions, of composite pipelines?
Or detailed versions of indiv algorithms?

Recommendation 3: favor pragmatic over grandiose (need achievable buy-in of data generation centers)

Hole: histology images not in Firehose (too big)

http://tcga.lbl.gov/biosig/tcgadownload.do

Morphometric data for integrative analysis can be downloaded at the Berkeley Cancer Morphometric Data wiki.

Biosig

TumorType
BLCA
BRCA
CESC
COAD
⋮
SKCM

Don’t Care? Level 1?

But a mapping could in principle be made.

Hole: custom analyses?

On Wiki?
In telecon slides?
On local disk of analyst?

How many such threads need to be tied together?

Recommendation 4: appoint reproducibility champion (morph data analysis/champion into this role, as freeze solidifies?)

Data/Analysis/Reproducibility Champion

Recommendation 5: create SKCM space in Synapse

Recommendation 6: write down phased plan
starting with sufficiently clear definitions & scope (Rec. 1, slide 4)
with clear but realistic dates/milestones