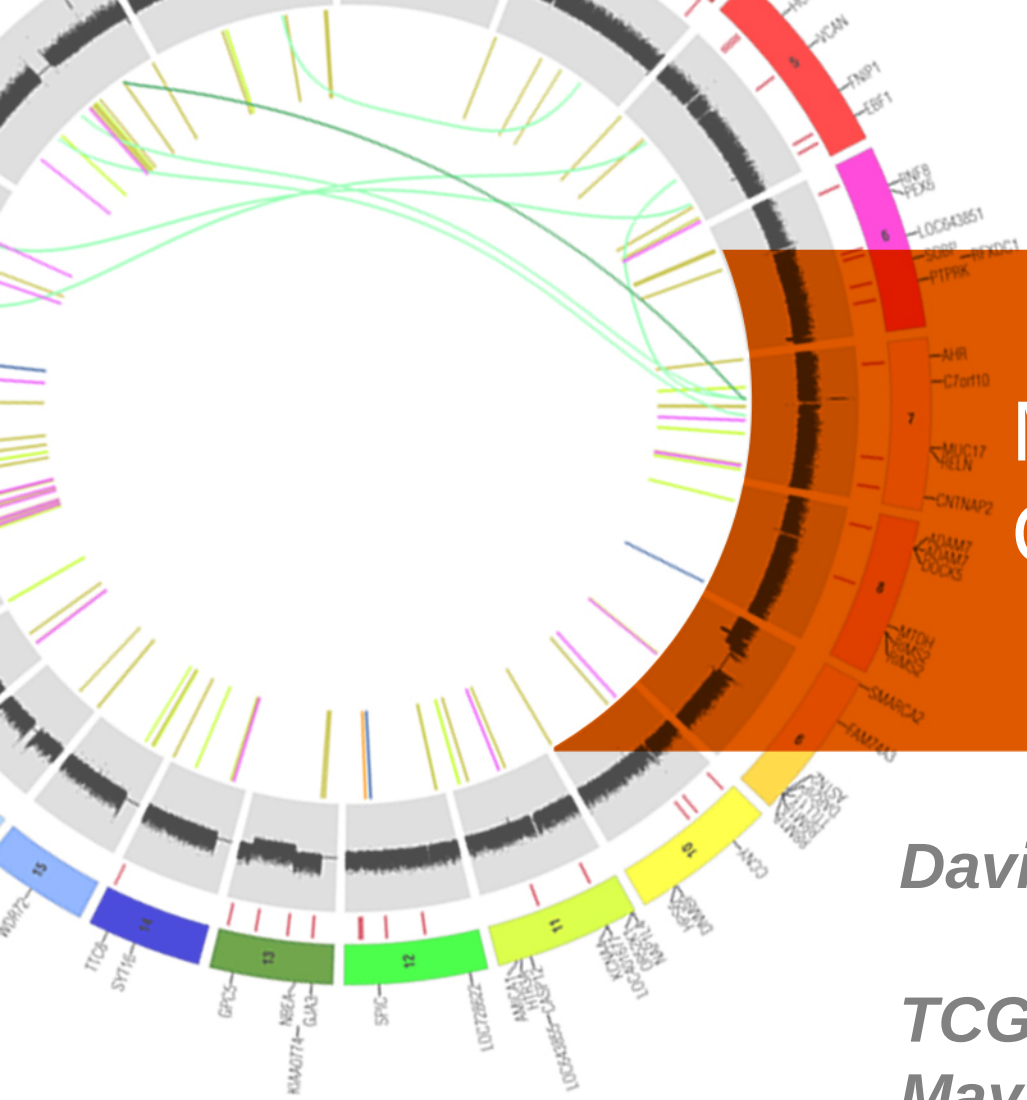


The Cancer Genome Atlas

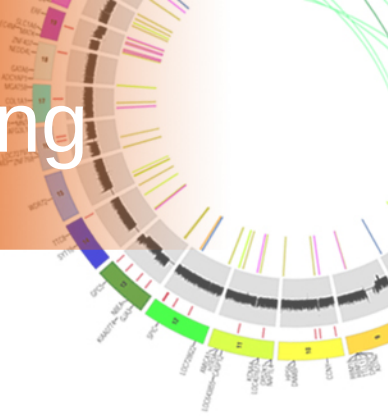
Multi-center Mutation Calling in TCGA

David A. Wheeler,

TCGA Symposium
May 13, 2014

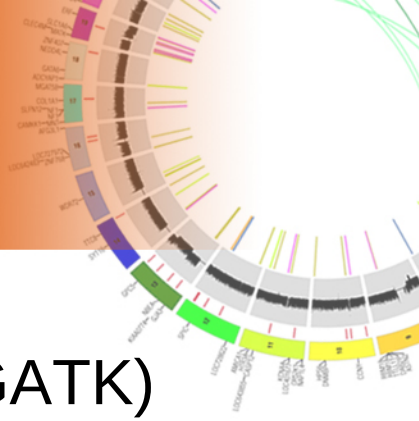
HumanGenomeSequencingCenter
BCM
Baylor College of Medicine

(The art of) Multi-center mutation calling



- Approaches to somatic mutation calling
- Early benchmarking of somatic mutation callers
- Early trials of 3-center calling and adoption of standards
- Current status of multi-center calling
- New developments in mutation calling

Sources of error in sequencing data



- Base calling error
  - Randomly distributed after Q recalibration (GATK)
  - Largely reflected in the Q-values
  - Estimated error rates ~10^{-3} per base (Illumina)
  - Filter most of it, except for calling in low allele fraction variants

- Mapping error and alignment ambiguities
  - Systematic
  - Depends on details of repeat structure of the genome
  - Repeat structure is different in tumor and normal
  - Depends on sequencing chemistry
  - Produces high-quality variation

Variant “truth engines” – 1st generation

- Atlas-SNP (BCM: Yu et al.)

$$\Pr(SNP \mid S_j, c)_j = \frac{\Pr(S_j \mid SNP, c) \times prior(SNP \mid c)}{\Pr(S_j \mid SNP, c) \times prior(SNP \mid c) + \Pr(S_j \mid error, c) \times prior(error \mid c)}$$

- MuTect (BI: Cibulskis et al.)

$$LOD_T = \log_{10}\left(\frac{L[M_f^m]}{L[M_0]}\right) \geq \log_{10}\delta - \log_{10}\left(\frac{P(m,f)}{(1-P(m,f))}\right) = \theta_T$$

- Pebbles (UCSC: Ma et al.)

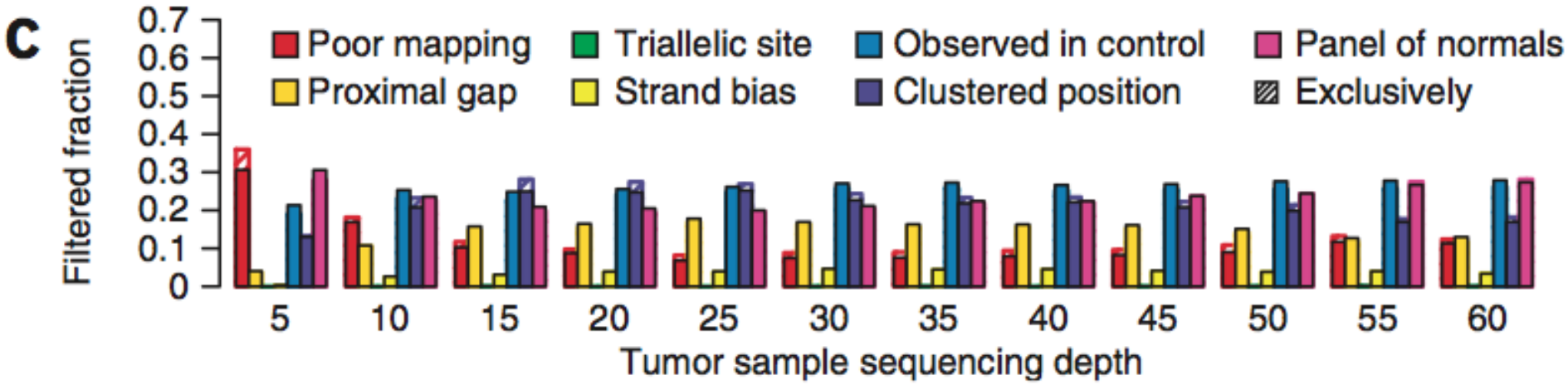
$$P(G_t, G_n) = \frac{n!}{n_A! n_T! n_G! n_C!} \prod P(d_i | G_T, G_N, c) \cdot \frac{n!}{n_A! n_T! n_G! n_C!} \prod P(d_i | G_n) \cdot P(G_n)$$

- BaSSoVac (TGI: Wendl & Ding)

$$P(R_{hi0}|\mathbf{S}) = \left(1 - \frac{4\,\varepsilon_{hi}}{3}\right)\frac{\phi_{1,ijk} + \phi_{2,ilmn}}{d_{il}} + \varepsilon_{hi} \qquad d_{il} = \phi_{3,l} + (\phi_{3,i} - \phi_{3,l})\,p_i \quad i \neq l.$$

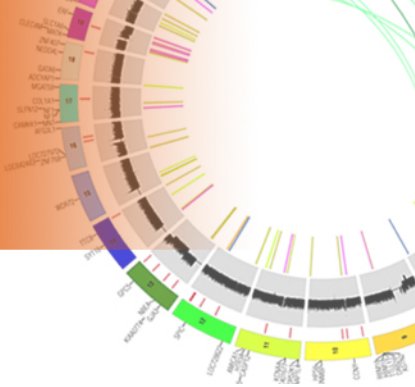
Downstream processing

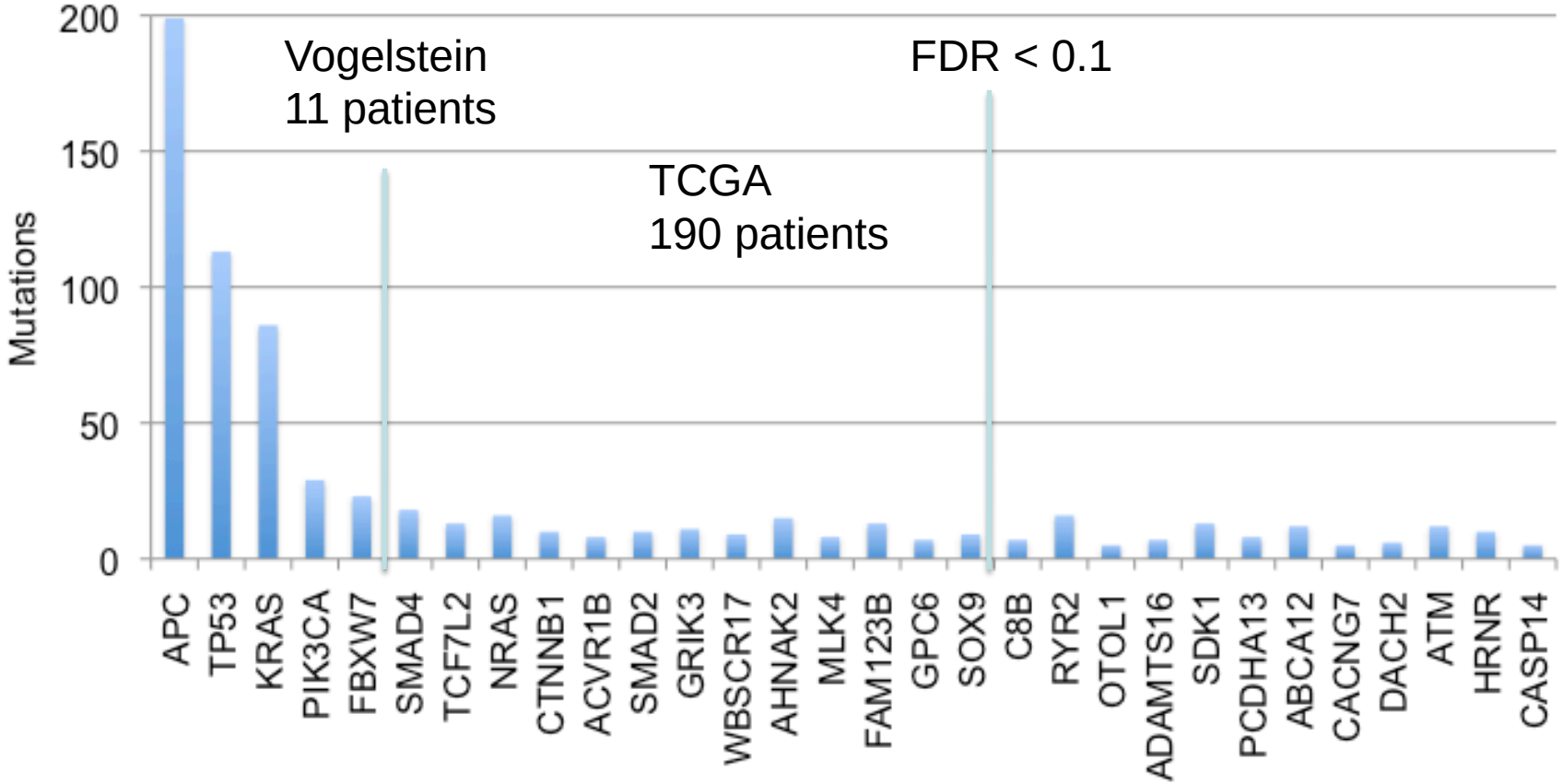
- Initial variation calls must be filtered by heuristic criteria
  - MuTect best documented 7 criteria

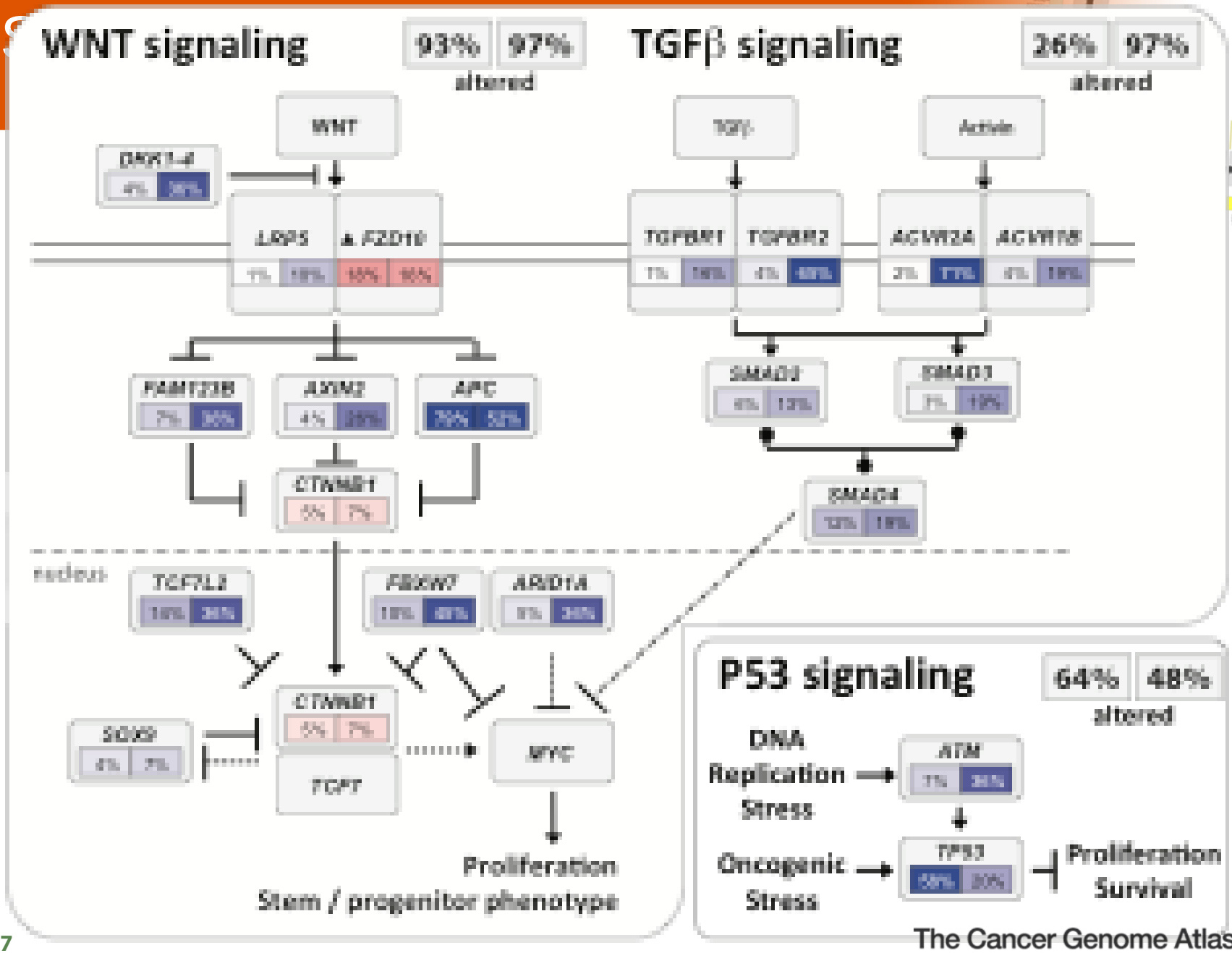


Cisbulskis et al. Nature Biotechnology 2013

TCGA Colorectal Cancer Genes

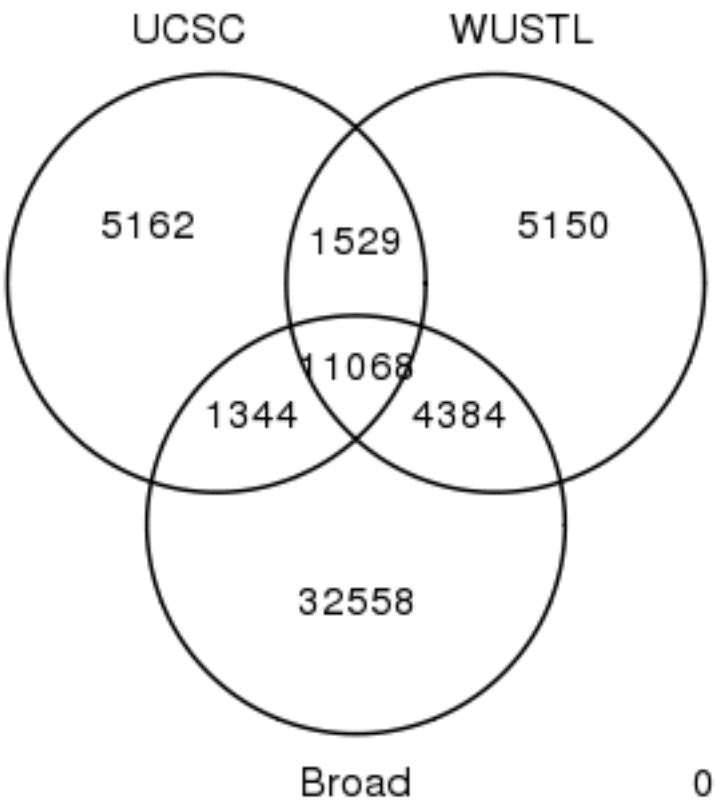




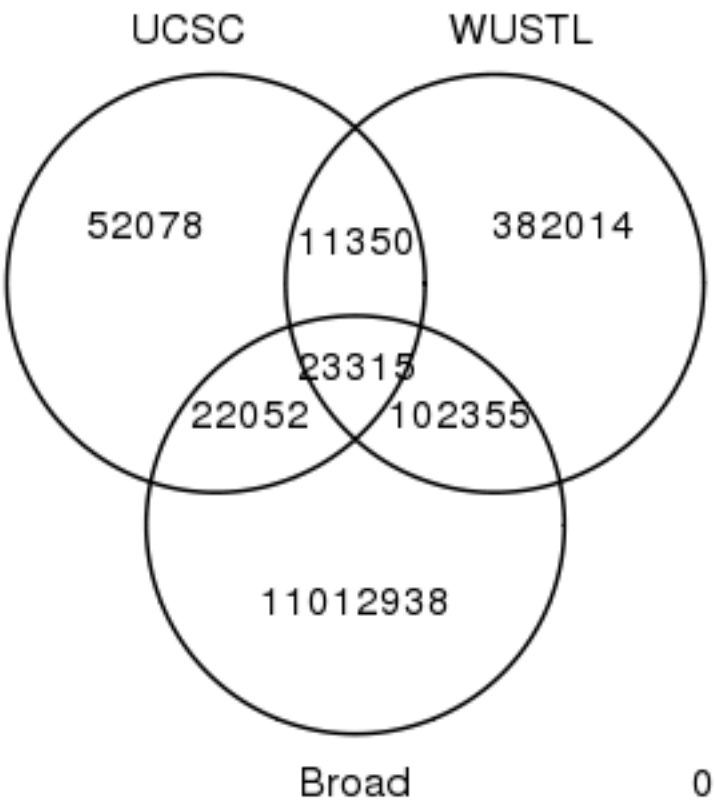

WNT signaling
93%
97%
altered
WNT
DKK1-4
LRP5
FZD10
FAM123B
AXIN2
APC
CTNNB1
TCF7L2
FBXW7
ARID1A
SOX9
CTNNB1
TCPT
MYC
Proliferation
Stem / progenitor phenotype
TGFβ signaling
26%
97%
altered
TGFβ
Activin
TGFBR1
TGFBR2
ACVR2A
ACVR1B
SMAD2
SMAD3
SMAD4
P53 signaling
64%
48%
altered
DNA Replication Stress
Oncogenic Stress
ATM
TP53
Proliferation
Survival

Benchmark 1 ca Feb 2011

Strict: Somatic Calls Passing Filters

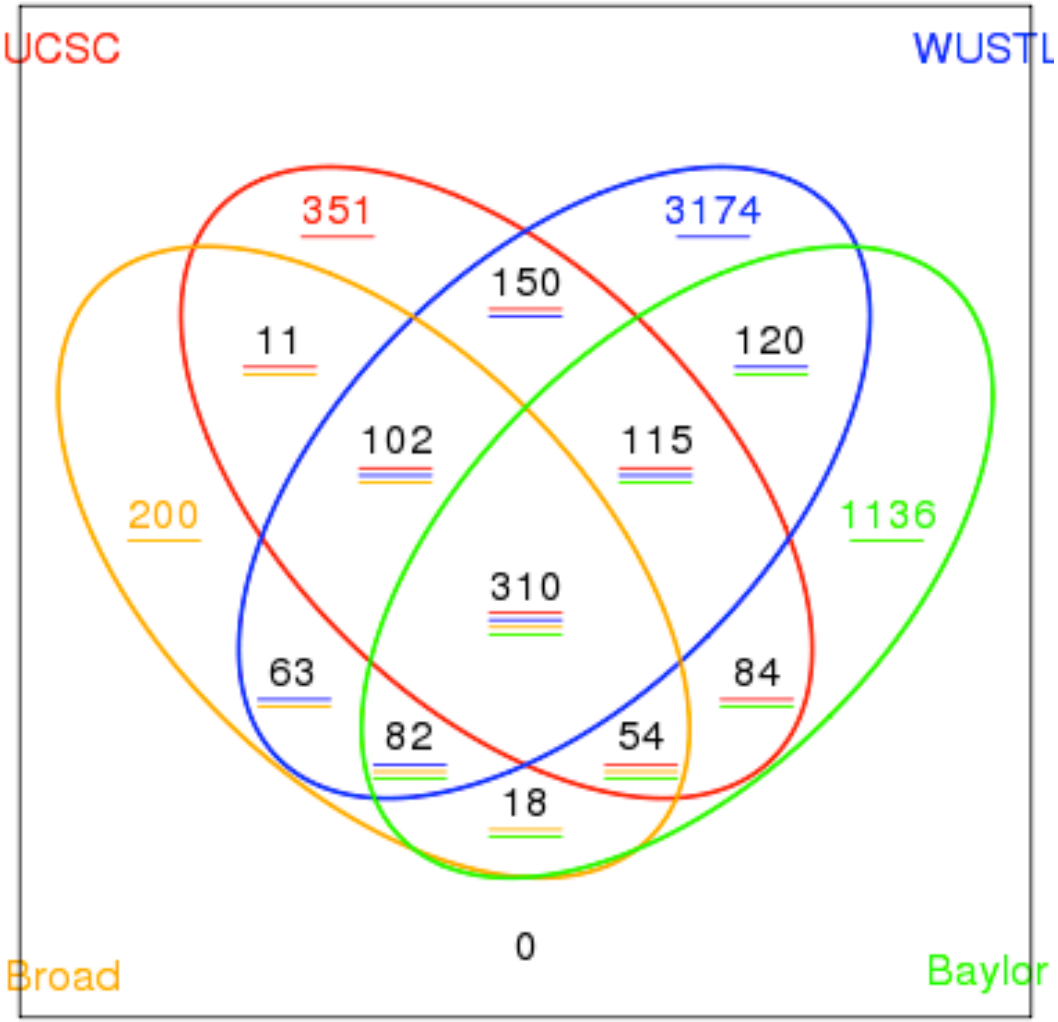


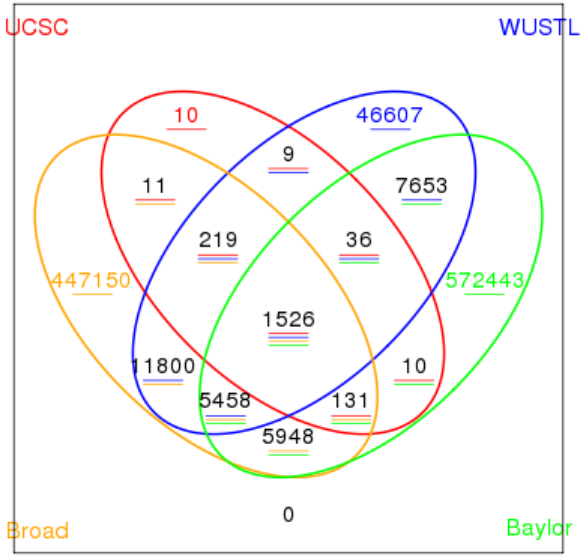
Loose: All Somatic Calls



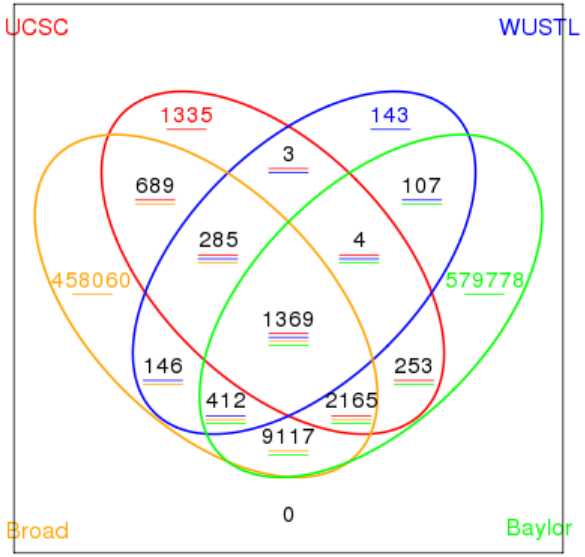
Benchmark 2 ca March 2011

Strict: Somatic Calls Passing Filters



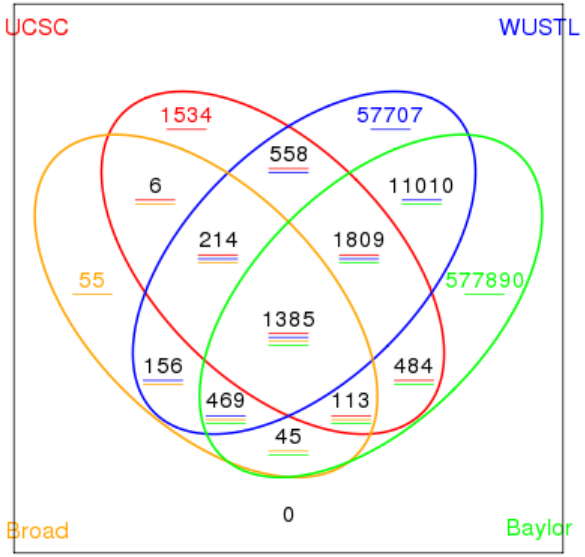


UCSC Top 100

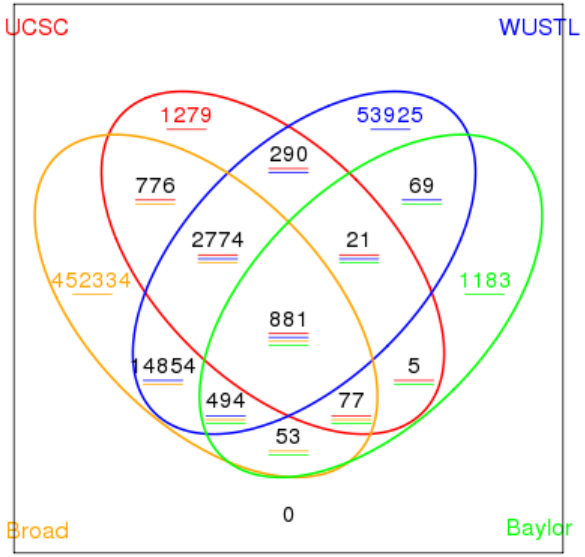


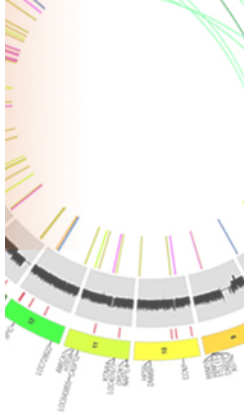
WUSTL Top 100

Broad Top 100



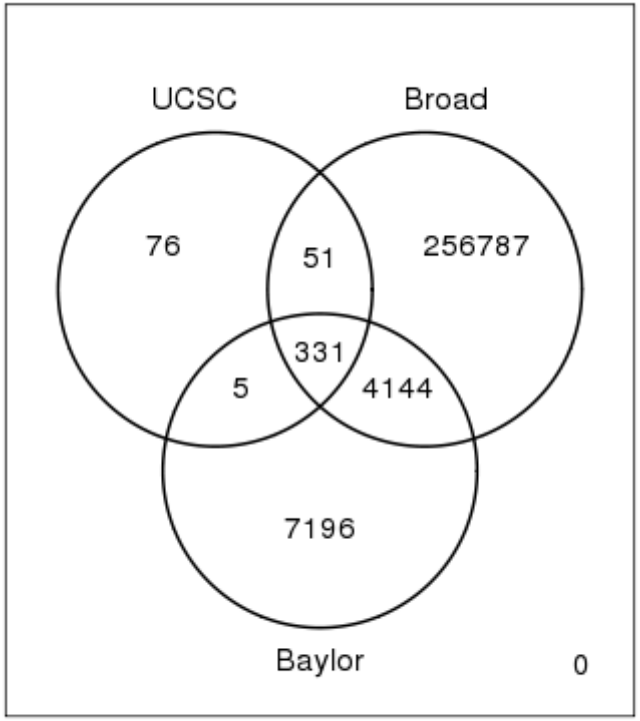
Baylor Top 100



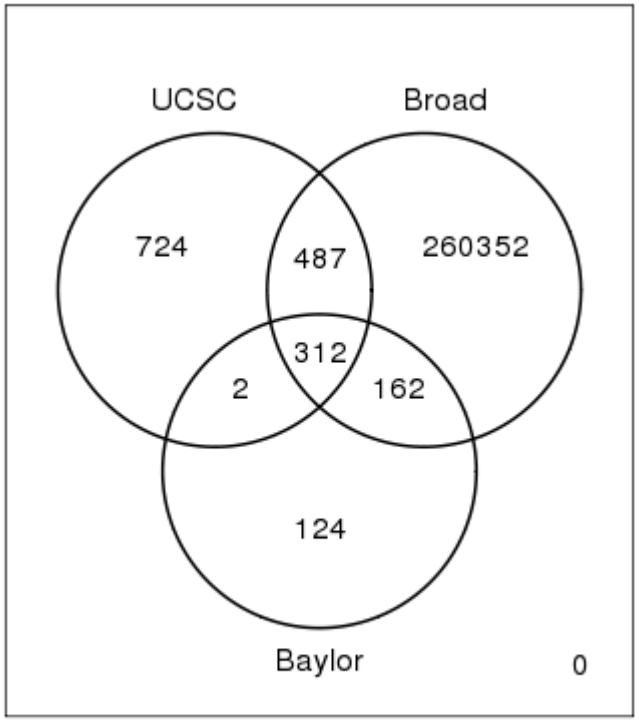


me Atlas

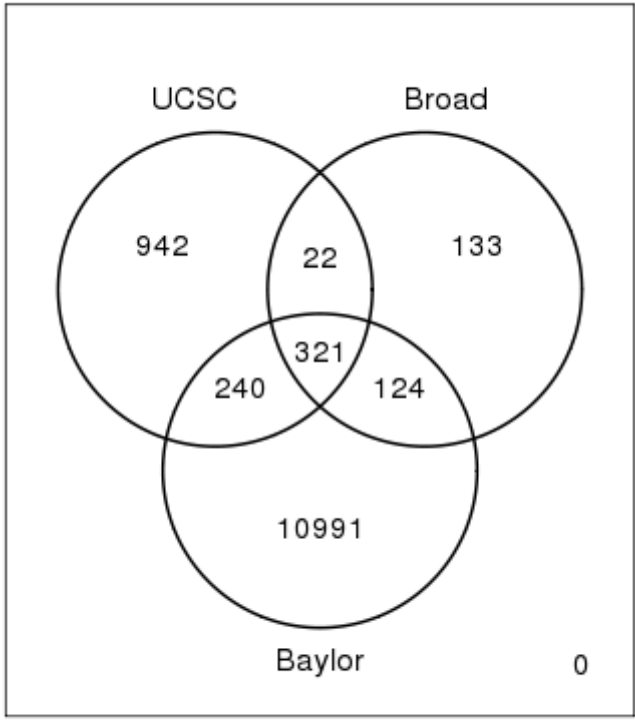
Colorectal adenocarcinoma



UCSC Top 100



Baylor Top 100

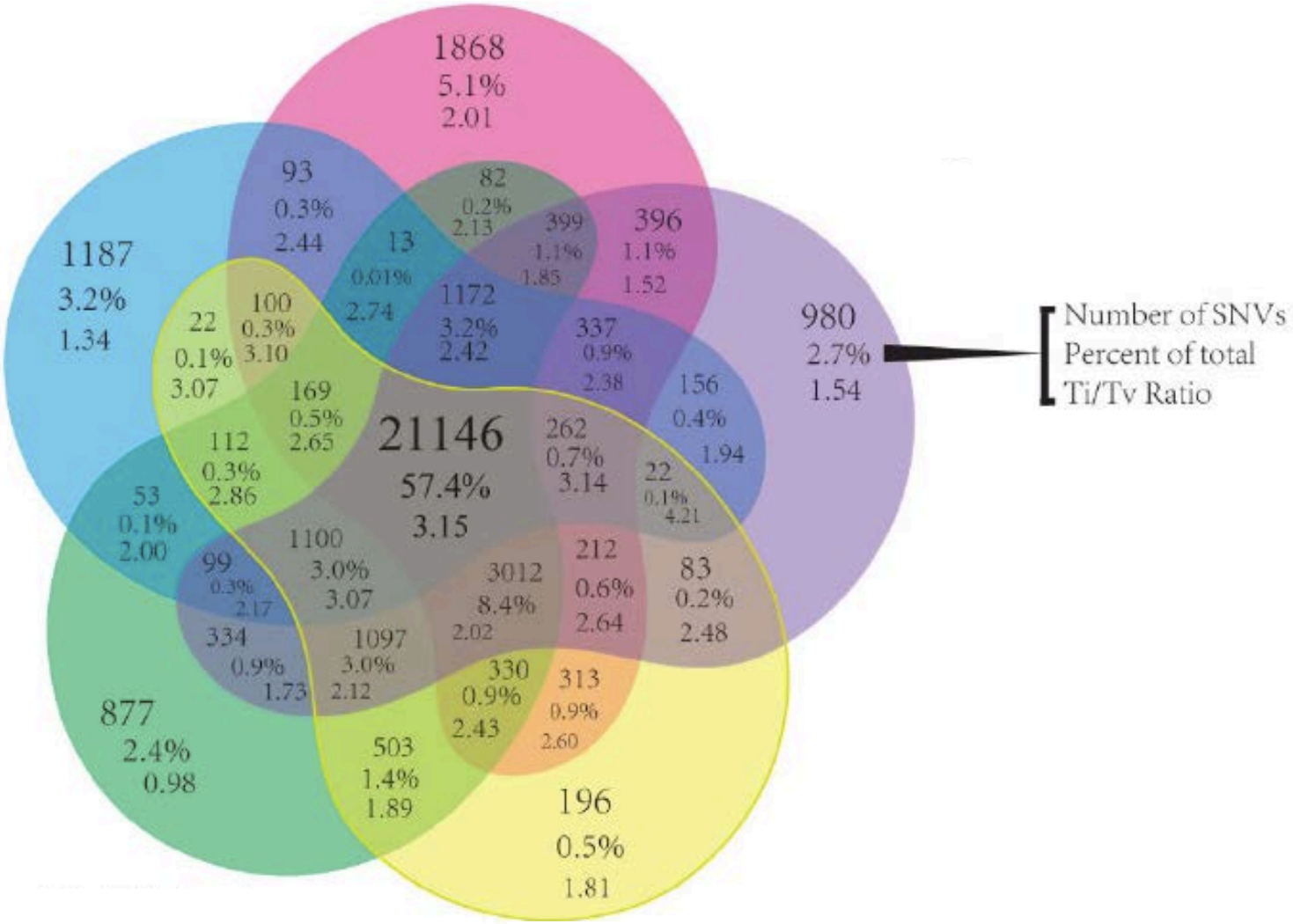


Broad Top 100

Early benchmarking conclusion

- Discordance between callers was high
- High quality calls, as defined by one caller were missed by other callers.
- Multi-center mutation calling may ameliorate these issues

Low concordance in (diploid) SNP calling

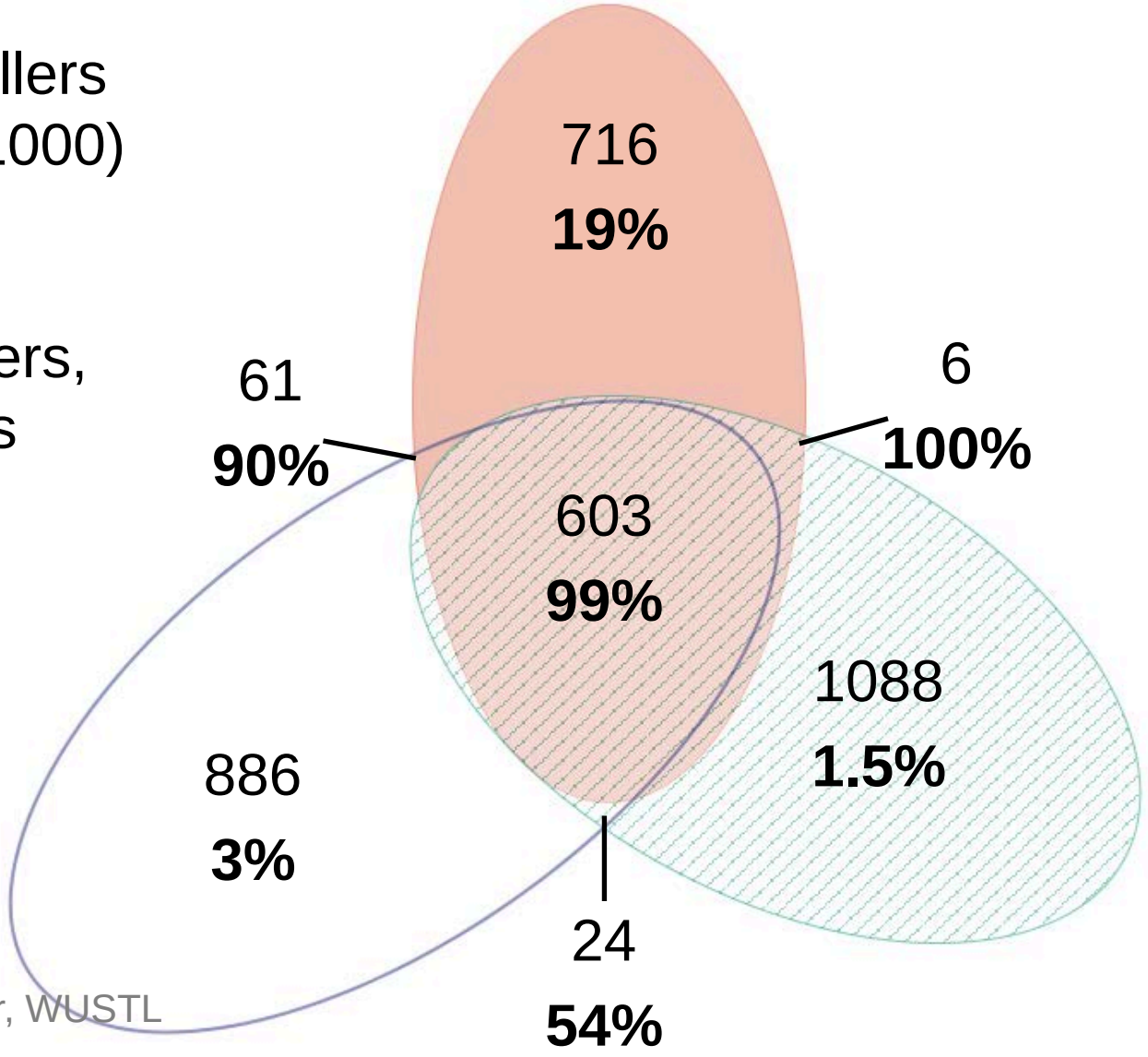


Acute Myeloid Leukemia

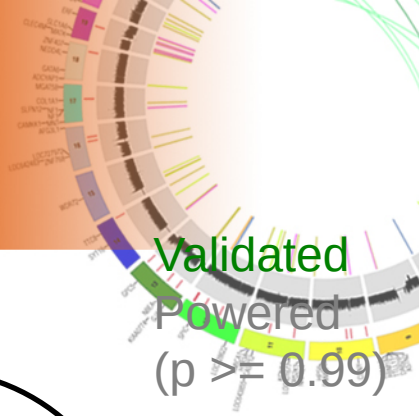
-- 3 somatic SNV callers
-- subset of calls (>1000) manually reviewed

- If we require 3 callers, ~253 false negatives

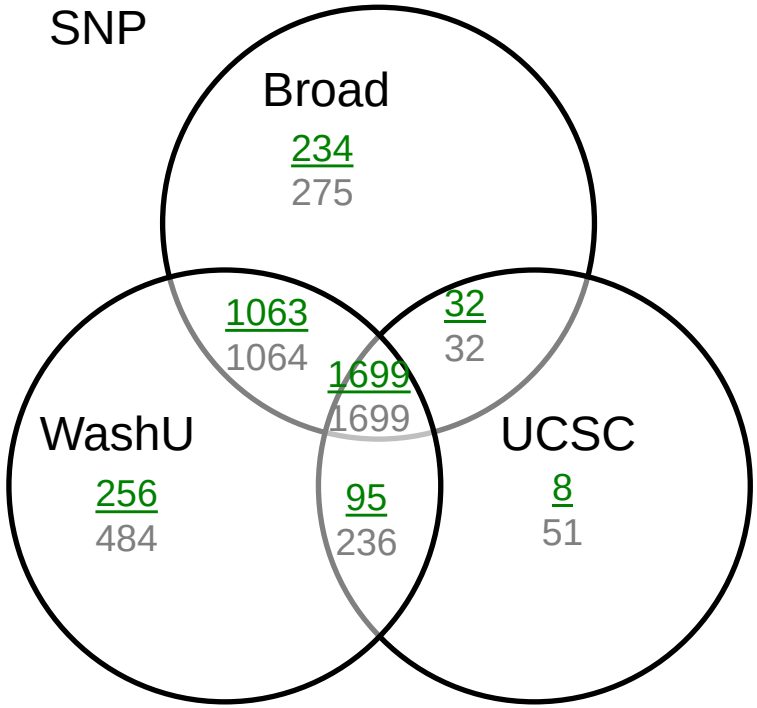
Number of calls
Validation percentage

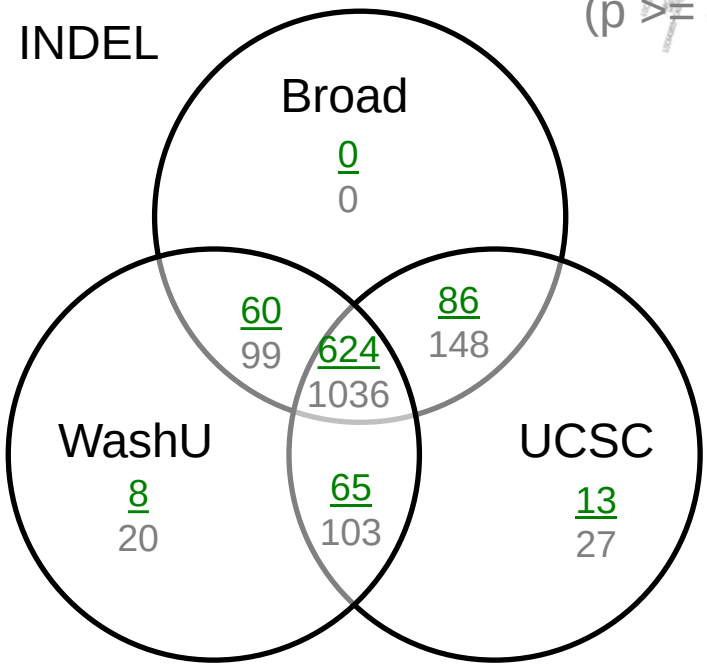


Lung Adenocarcinoma



161 patients





Caller Validation (SNPs):

Broad: 98.6%

UCSC: 90.9%

WashU: 89.4%

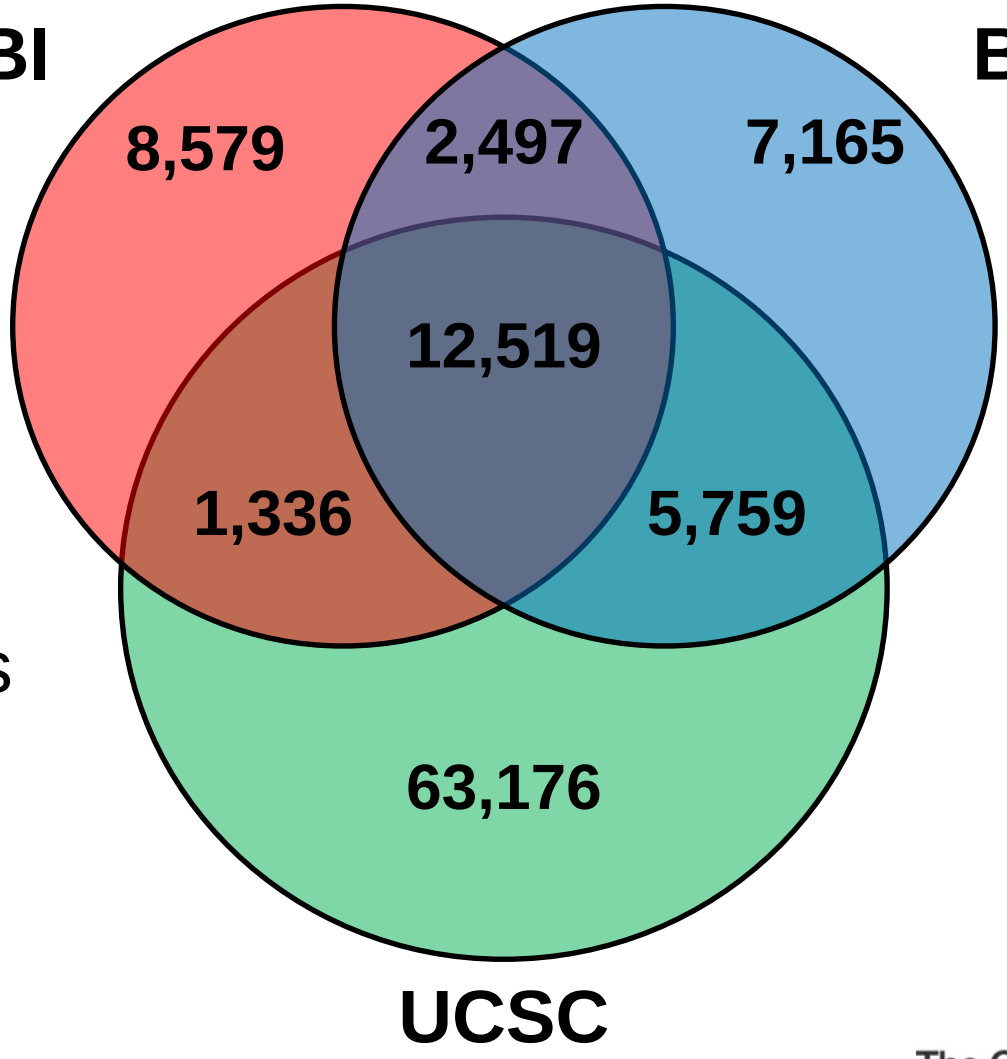
Caller Validation (Indels):

Broad: 60%

WashU: 60%

UCSC: 60%

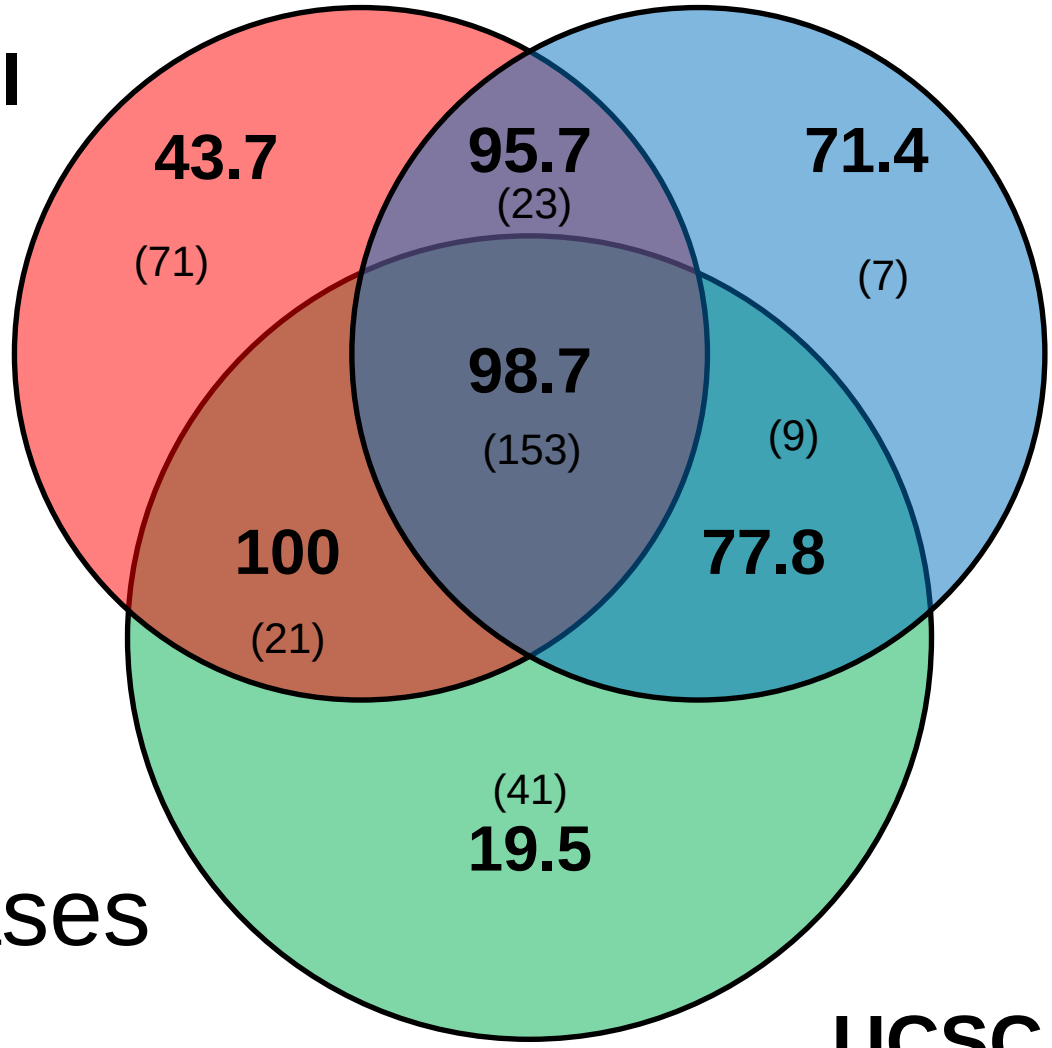
Somatic mutations called by 3 centers
(including Illumina and Solid, BI did not call on Solid)



506 patients

Indels and SNVs

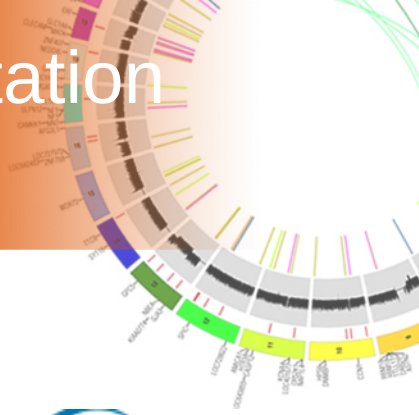
Validation of 325 Illumina SMG SNV on Ion Torrent



177 cases

The Cancer Genome Atlas

Meta caller developed based on multiple mutation callers calibrated by validation data



Kim and Speed *BMC Bioinformatics* 2013, **14**:189
http://www.biomedcentral.com/1471-2105/14/189

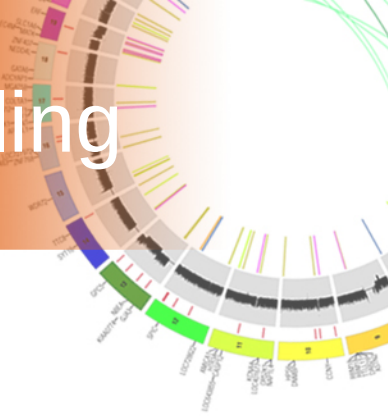
BMC Bioinformatics

RESEARCH ARTICLE Open Access

Comparing somatic mutation-callers: beyond Venn diagrams

Su Yeon Kim[1*] and Terence P Speed[1,2*]

2012-3 formalization of multicenter calling

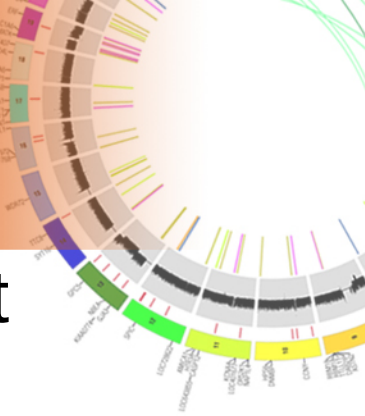


- Mutations callers are improving
- Different callers detect different events
- Validation cycles take too long and cause delay in submissions
- 3-center calling
  - Multiple caller stratify the calls into high to low quality
  - Initial time line allowed 6 weeks (now 3 weeks)

Marker paper submission

- MAF contains 3 callers with annotation of which group provided each call. (column 3 of the MAF)
- Significantly Mutated Gene list (e.g. MutSig, MUSIC) uses calls supported by two-centers (which guarantee high accuracy).
- The resulting SMGs can be used for submission.

Final publication includes validation

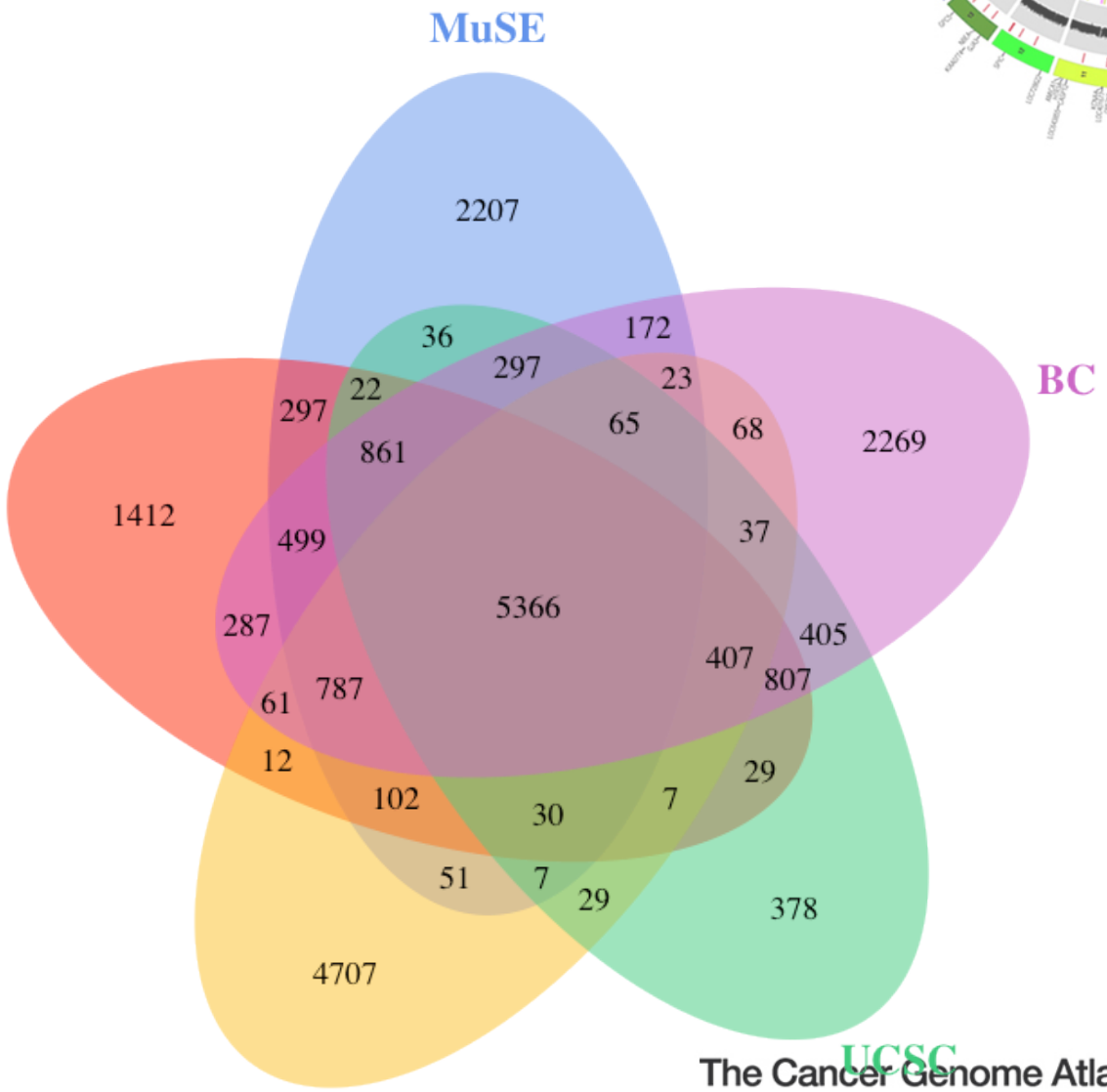


- Validation requires a second independent sequencing event.
- Visual inspection of reads (e.g., IGV) is not validation; however, visual inspection improves accuracy. MAF was designed to capture that in the “Verification” column.
- Validation results are captured in separate BAM files and submitted.
- Validation should include mutations in genes on SMG, and include mutations found by single caller.

Adrenal cortical carcinoma 91 patients: SNVs

New callers can be easily added, and therefore have role in marker paper.

With 5 callers require 3/5 being tested.



The Cancer Genome Atlas

Second Generation Mutation Callers

- Increased sophistication in heuristic filters
- Increased sophistication in underlying genetic models

(poster #60)

Evolutionary Distance (d) $\propto$
Substitution Rate (μ) $\times$ Tumor Development Time (t)

"To" Tumor Sequence

"From" Reference Sequence	A	C	G	T
A	—	π_C	$\pi_G\kappa$	π_T
C	π_A	—	π_G	$\pi_T\kappa$
G	$\pi_A\kappa$	π_C	—	π_T
T	π_A	$\pi_C\kappa$	π_G	—

μ

π : allele fraction
κ : transition / transversion rate ratio
μ : scaled substitution rate
t : tumor development time

- MuSE – HGSC/MDACC
  - Distance measure per position per sample reflecting mutation evolution
  - Uncertainty estimates based on Bayesian Markov model
- Viper – Wash U
- MuTect v 2

Dream Challenge

Synapse CONTRIBUTE to the CURE

Search Forum Register Login

ICGC-TCGA DREAM Mutation Calling challenge

Sharing

Synapse ID: syn312572

Single Nucleotide Variants

ID	Submission Name	Team	Status	SMC Challenge Eligible	Number of Calls (Censor:Chr1)	Number of Calls (All)	Sensitivity (Censor:Chr1)	Sensitivity (All)	Specificity (Censor:Chr1)	Specificity (All)	Balanced Accuracy (Censor:
2343085	MuTect - L10	Broad SMC	SCORED	YES	3225	3476	0.96708	0.96719	0.98388	0.98360	0.97548
2368537	LoFreq Somatic: beta-3-159 dbac	LoFreq Somatic - GIS	SCORED	NO	3179	3422	0.95093	0.94908	0.98144	0.98042	0.96619
2346677	nRex3 mpileup-gatk-ext	DellyTeam	SCORED	YES	3235	3489	0.95672	0.95757	0.97032	0.97019	0.96352
2350981	nRex4 mpileup-gatk-ext	DellyTeam	SCORED	YES	3340	3606	0.96586	0.96747	0.94880	0.94842	0.95733
2350983	nRex5 mpileup-gatk-ext	DellyTeam	SCORED	YES	3338	3604	0.96556	0.96719	0.94907	0.94867	0.95732
2367742	LoFreq Somatic: beta-3-159 dba	LoFreq Somatic - GIS	SCORED	NO	3119	3360	0.93386	0.93267	0.98237	0.98125	0.95811
2346241	nRex mpileup-gatk-ext	DellyTeam	SCORED	YES	3341	3600	0.96525	0.96549	0.94792	0.94806	0.95659
2350992	dmut0.set1.02_5	DMUT	SCORED	YES	3189	3442	0.94240	0.94286	0.96958	0.96833	0.95599
2350994	dmut0.set1.02_6	DMUT	SCORED	YES	3191	3444	0.94240	0.94286	0.96898	0.96777	0.95569
2350974	dmut0.set1.02_4	DMUT	SCORED	YES	3201	3454	0.94240	0.94286	0.96595	0.96497	0.95417
2367150	LoFreq Somatic: beta-3-159 db	LoFreq Somatic - GIS	SCORED	NO	3231	3482	0.94727	0.94569	0.96193	0.96008	0.95460
2347517	dmut0.set1.01_4	DMUT	SCORED	YES	3212	3464	0.94240	0.94286	0.96264	0.96218	0.95252
2367520	LoFreq Somatic: beta-3-159 b100 dba	LoFreq Somatic - GIS	SCORED	NO	3190	3435	0.93904	0.93833	0.96583	0.96565	0.95244
2367152	LoFreq Somatic: beta-3-159 db no-p1	LoFreq Somatic - GIS	SCORED	NO	3252	3503	0.94910	0.94738	0.95756	0.95604	0.95333
2350872	dmut0.set1.001_4	DMUT	SCORED	YES	3264	3523	0.94697	0.94795	0.95190	0.95118	0.94943
2347464	Dream_Set1_MuSE_Setting8	Wang-Wheeler-HGSC	SCORED	YES	3086	3329	0.91984	0.92079	0.97797	0.97777	0.94890

Synapse CONTRIBUTE to the CURE

Search Forum Register Login

ICGC-TCGA DREAM Mutation Calling challenge

Sharing

Synapse ID: syn312572

Wiki Files

Single Nucleotide Variants

ID	Submission Name	Team	Status	SMC Challenge Eligible	Number of Calls (Censor:Chr1)	Sensitivity (Censor:Chr1)	Specificity (Censor:Chr1)	Balanced Accuracy (Censor:Chr1)
2363577	MuTect - L10F	Broad SMC	SCORED	YES	3850	0.96054	0.99273	0.97664
2385689	Dream_Set2_MuSE_Setting3	Wang-Wheeler-HGSC	SCORED	YES	3932	0.96004	0.97152	0.96578
2368737	Dream_Set2_MuSE_Setting2	Wang-Wheeler-HGSC	SCORED	YES	3805	0.94370	0.98686	0.96528
2375120	nRex	DellyTeam	SCORED	YES	3897	0.95124	0.97126	0.96125
2368734	Dream_Set2_MuSE_Setting1	Wang-Wheeler-HGSC	SCORED	YES	3844	0.94396	0.97711	0.96053
2367176	mutect_conta	SLC_platform	SCORED	YES	3911	0.94999	0.96650	0.95825
2367734	nRex mpileup-gatk-exta	DellyTeam	SCORED	YES	4006	0.95979	0.95332	0.95655
2400705	LoFreq Somatic: beta-4-50-g2ce040e t-a5 n-a3p	LoFreq Somatic - GIS	SCORED	NO	3801	0.93365	0.97737	0.95551

Synapse CONTRIBUTE to the CURE

Search Forum Register Login

ICGC-TCGA DREAM Mutation Calling challenge

Sharing

Synapse ID: syn312572

Single Nucleotide Variants

ID	Submission Name	Team	Status	SMC Challenge Eligible	Number of Calls (Censor:ChrMask)	Sensitivity (Censor:ChrMask)	Specificity (Censor:ChrMask)	Balanced Accuracy (Censor:ChrMask)
2470044	viper.v4	WashU	SCORED	YES	5876	0.90987	0.98962	0.94975
2468117	viper.v3d	WashU	SCORED	YES	5904	0.91144	0.98662	0.94903
2453885	Dream_Set3_MuSE_Setting4	Wang-Wheeler-HGSC	SCORED	YES	6050	0.92270	0.97471	0.94871
2456287	MuTectL700W	Broad SMC	SCORED	YES	5942	0.91410	0.98317	0.94863
2456202	MuTectL750W	Broad SMC	SCORED	YES	5903	0.91066	0.98594	0.94830
2463211	Dream_Set3_MuSE_Setting6	Wang-Wheeler-HGSC	SCORED	YES	5996	0.91738	0.97782	0.94760
2460162	viper.v2	WashU	SCORED	YES	5836	0.90455	0.99058	0.94756
2467165	viper.v3c	WashU	SCORED	YES	5918	0.91034	0.98310	0.94672
2460633	Dream_Set3_MuSE_Setting5	Wang-Wheeler-HGSC	SCORED	YES	5926	0.91097	0.98245	0.94671
2470029	varScan.snv.v7c	WashU	SCORED	YES	5819	0.90189	0.99055	0.94622
2463247	Dream_Set3_MuSE_Setting8	Wang-Wheeler-HGSC	SCORED	YES	6156	0.92787	0.96329	0.94558
2453883	Dream_Set3_MuSE_Setting3	Wang-Wheeler-HGSC	SCORED	YES	5836	0.90174	0.98749	0.94461
2420793	mutect_noise	SLC_platform	SCORED	YES	5748	0.89407	0.99408	0.94408

Conclusions

- TCGA paradigm mutation discovery is improved by multicenter calling
  - Decreased FN rates
  - Delivers a set of somatic SNVs of calibrated accuracy
  - Accelerates submission of marker papers
  - Stimulates development of new mutation callers by providing ‘benchmarking’ on the fly.
  - A formal “meta-caller” was developed which may be useful in retrospectively refining mutation calls from TCGA tumor sets
- Multi-center mutation calling has not been applied to other mutation modalities. Needs to be tested.

Acknowledgments

- Wash U
  - Li Ding
  - Chris Miller
  - Mike McClellan
  - Cyriac Kandoth
- Broad
  - Gaddy Getz
  - Mara Rosenberg
  - Kris Cisbulskis
- MD Anderson
  - Wenyi Wang
  - Yu Fan
- BCM
  - Liu Xi
  - Caleb Davis

Clear cell renal cell carcinoma (KIRK)

- >500 patients