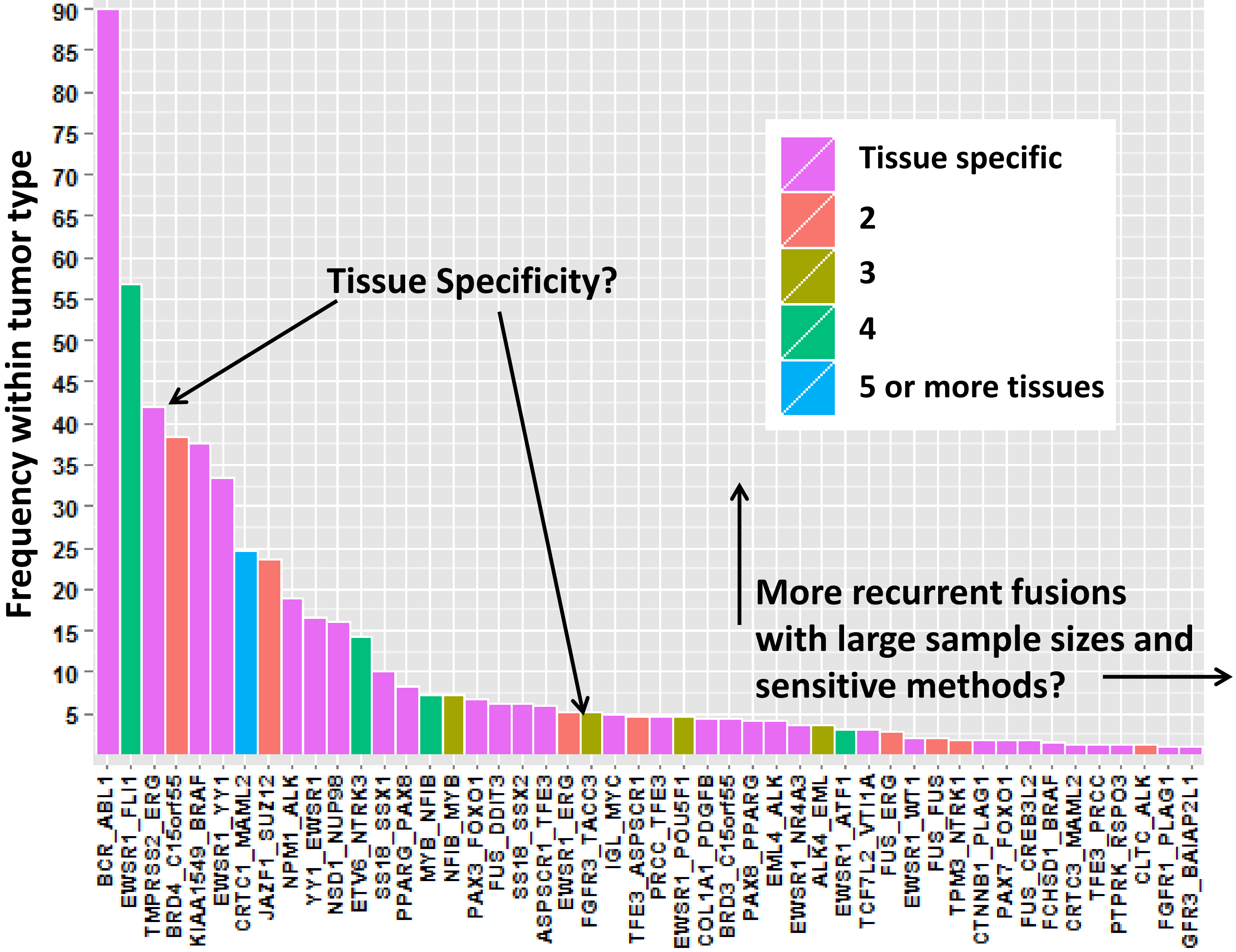
Discovery and functional characterization of recurrent gene fusions across 7,175 primary tumors

Chai Bandlamudi

Kevin White Lab
Institute for Genomics and Systems Biology
University of Chicago

Known recurrent fusions (COSMIC)

Only 48 recurrent fusions in COSMIC with frequency 1% or higher

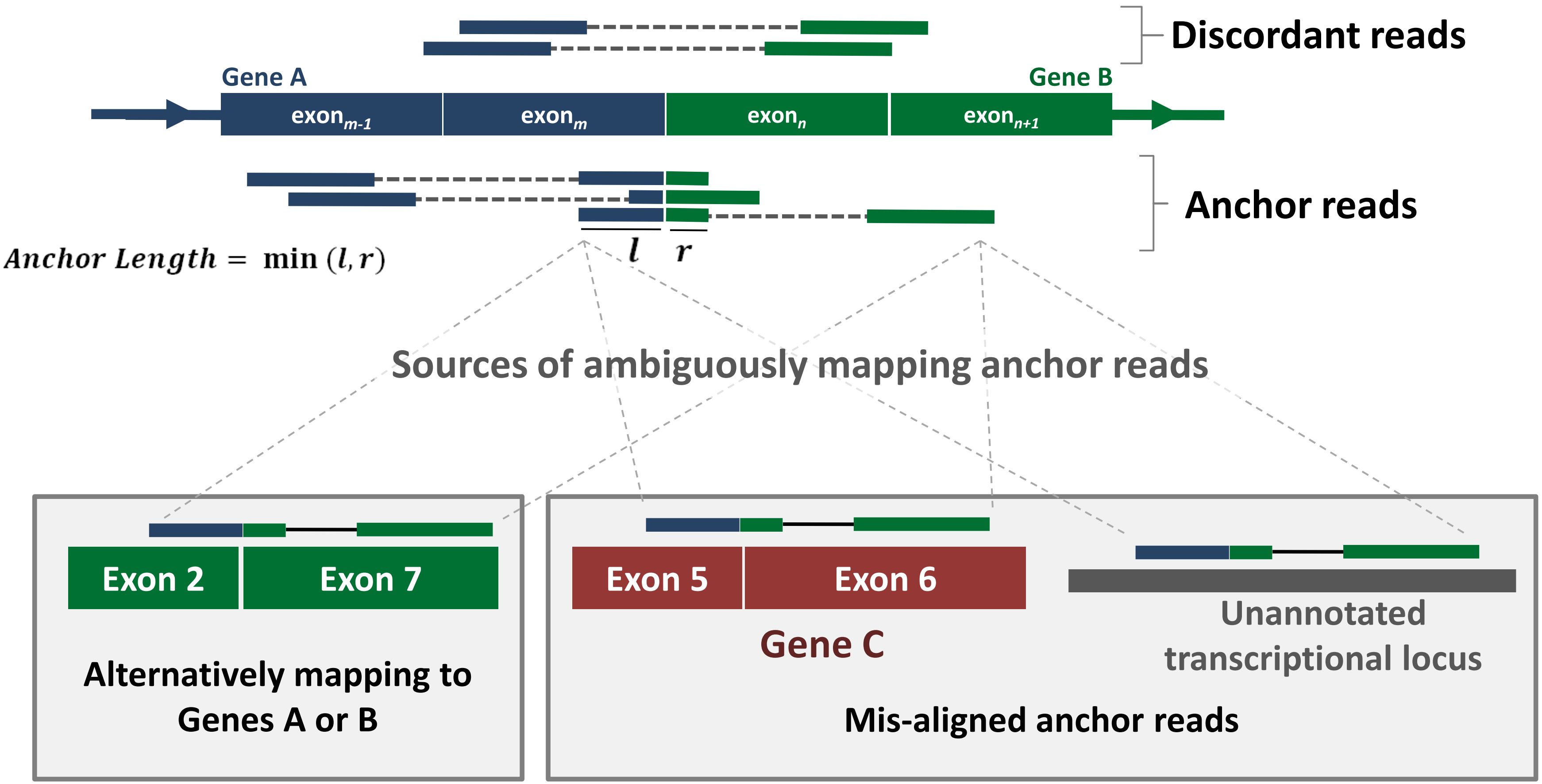


Tissue Specificity?

More recurrent fusions with large sample sizes and sensitive methods?

Compiled from fusions in COSMIC database (v66)

Fusion discovery from transcriptome sequencing



Discordant reads

Gene A

Gene B

exon_{m-1} exon_{m} exon_{n} exon_{n+1}

Anchor reads

$$\mathit{Anchor\ Length} = \min(l, r)$$

l r

Sources of ambiguously mapping anchor reads

Exon 2 Exon 7

Alternatively mapping to Genes A or B

Exon 5 Exon 6

Gene C

Unannotated transcriptional locus

Mis-aligned anchor reads

Minimum Overlap Junction Optimizer (MOJO)

Method to identify <u>canonical</u> fusions from paired-end transcriptome data

b. Identify candidate fusions

All paired-end reads

Align to transcriptome

Unmapped

trim 5' ends

Unmapped reads

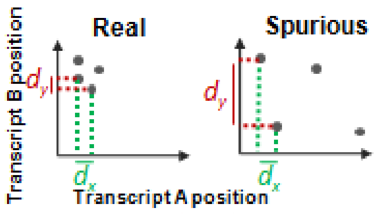
Align to genome

Filter discordant reads

- Remove PCR duplicates
- Stringency for repetitive regions
- Homology between genes

Construct discordant clusters

Filter library artifacts



Criteria: $\max(d_x, d_y) < 500bp$

Candidate fusion genes

c. Identify anchor reads

Identify putative junction exons

Construct fusion junction library

Align unmapped reads

Predicted Anchor Reads

d. Filter anchor reads

<u>Technical artifacts</u> - mis-aligned anchor reads

Iterative search

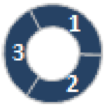
Alternative mapping to genome

Alternative mapping to transcriptome

Mapping to novel splice junctions

<u>Biological artifacts:</u> spurious anchors

Intra-genic rearrangements



Back-splicing

Anti-sense transcription

e. Filter fusion junctions

Alignment or annotation artifacts

fusion junctions with split-reads peak more than 2-SDs from fusion junction are filtered out

Random chimeras between highly expressed genes

$$RC_f = \frac{DR_f}{\Sigma_j AR_{fj}}; \quad require\ RC_f < \delta$$

DR_f - # of unique discordant reads supporting fusion f

AR_{fj} - # of anchor reads supporting junction j of fusion f

RC_f - Random chimera ratio for fusion f

δ - cut-off above which fusion f will be filtered out

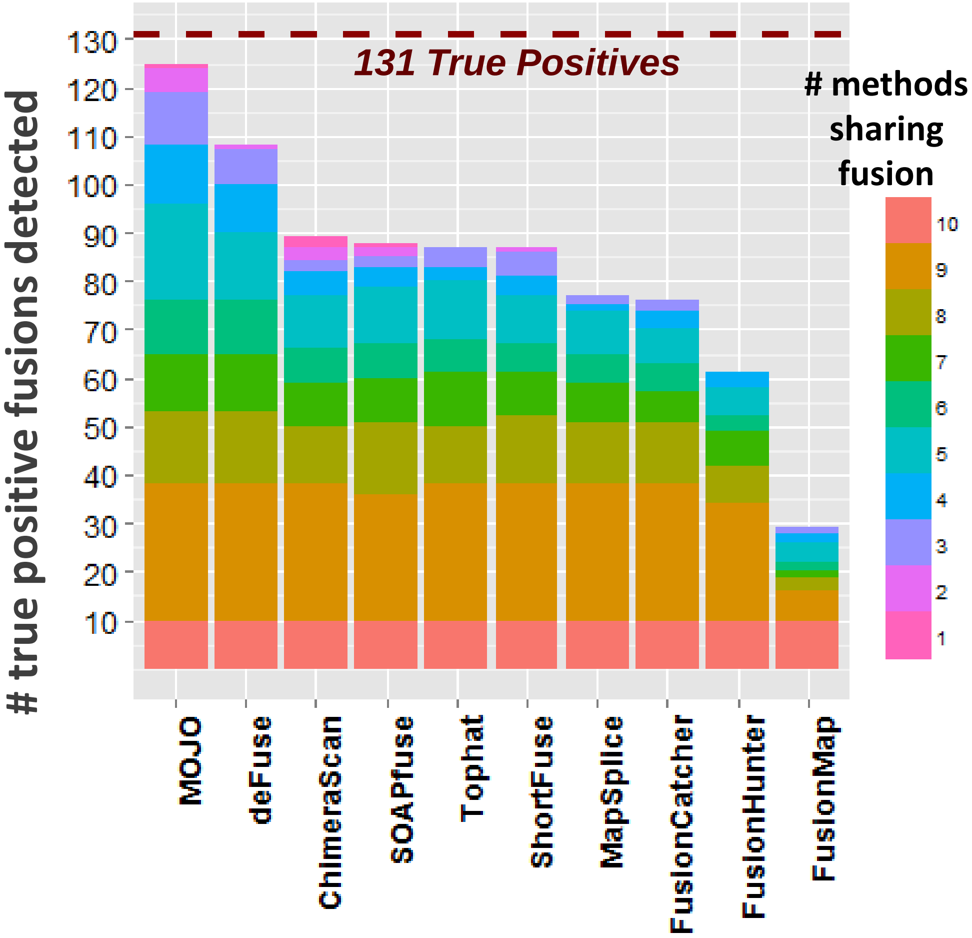
e. Scoring criteria

- d_s - # of discordant reads
- A_n - # of anchor reads with anchor length 'n'
- PA_n - # of proper anchor reads with uniquely mapped 'other-read' ends

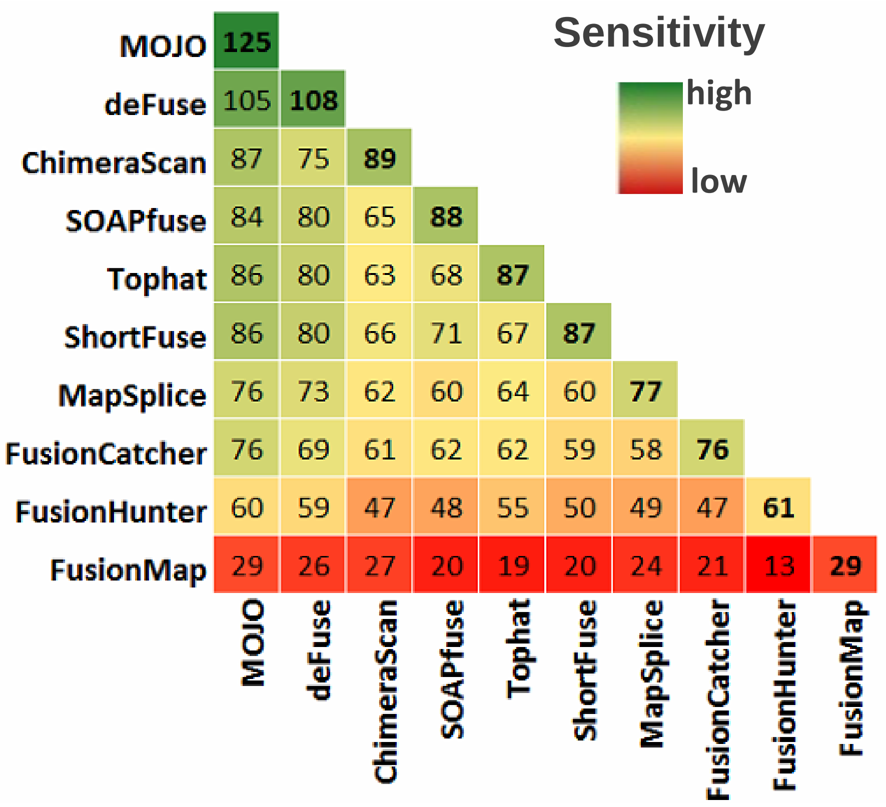
Bandlamudi et al., in preparation

MOJO Sensitivity

Comparison dataset: 18 cell lines transcriptomes with previously validated fusions



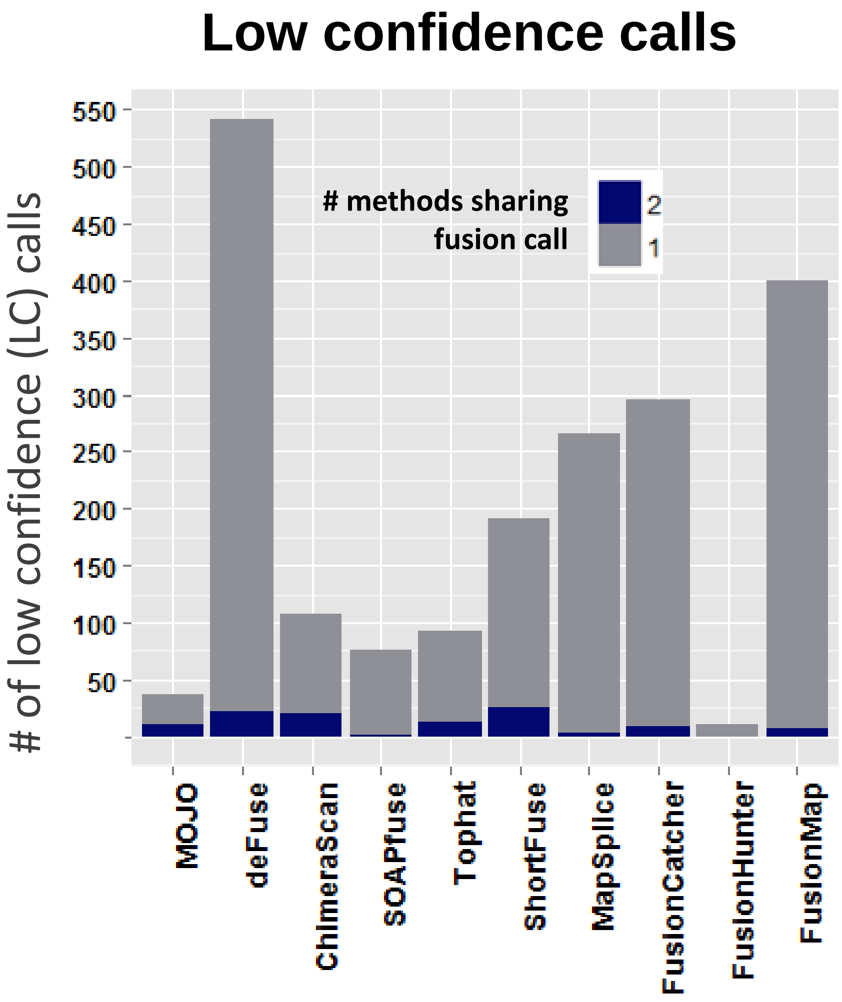
Sensitivity (canonical fusions)

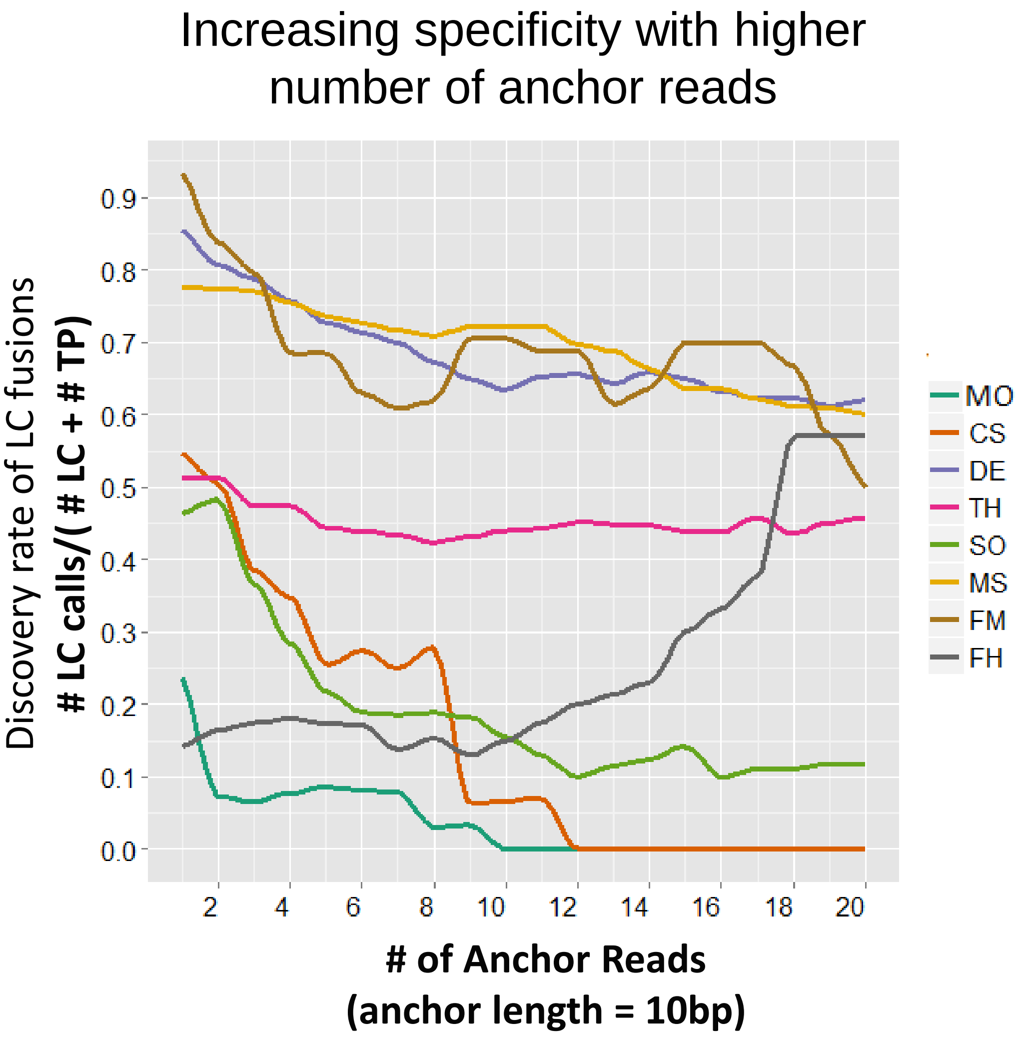


Overlap in sensitivity (# of fusions)

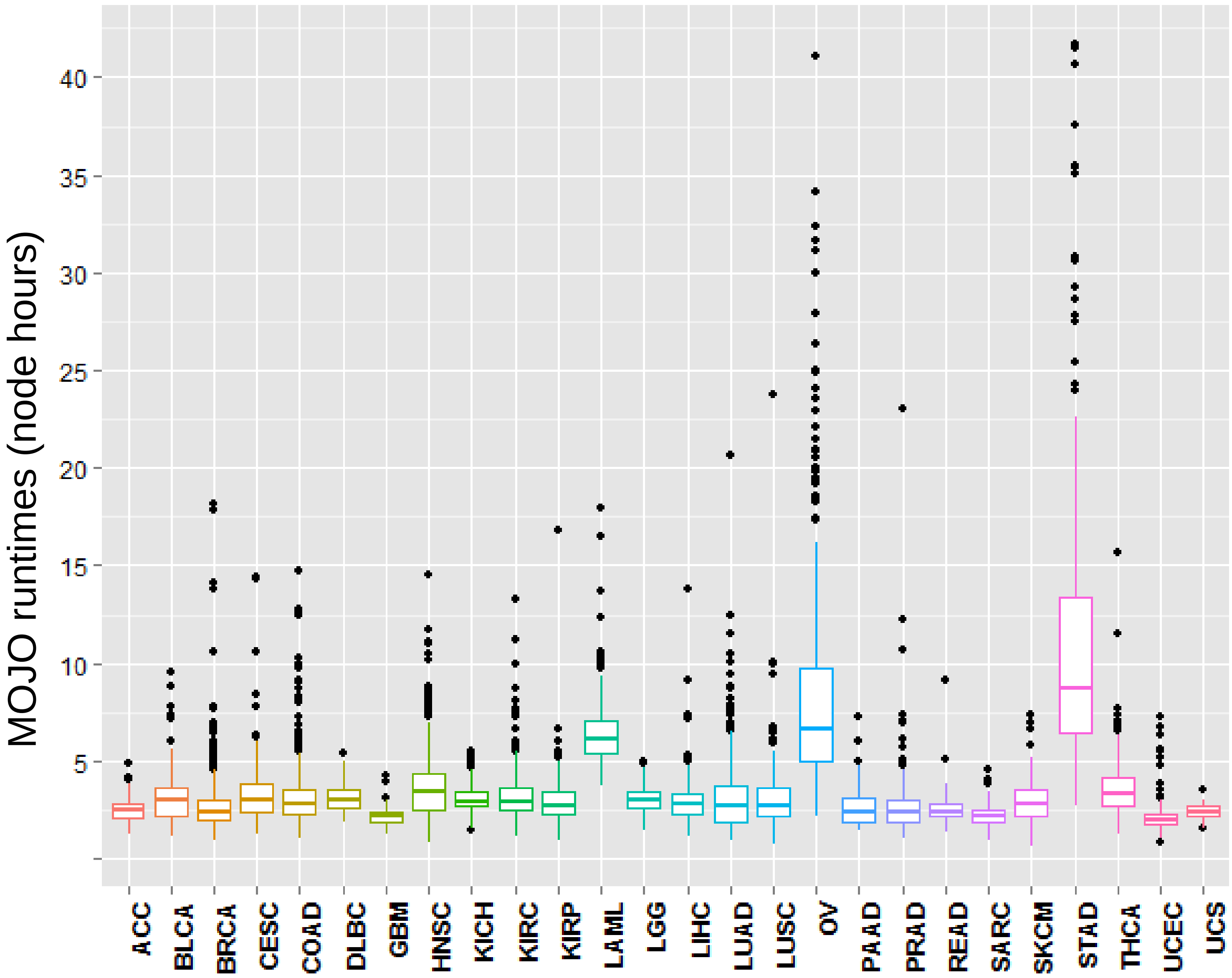
	MOJO	deFuse	ChimeraScan	SOAPfuse	Tophat	ShortFuse	MapSplice	FusionCatcher	FusionHunter	FusionMap
MOJO	**125**									
deFuse	105	**108**								
ChimeraScan	87	75	**89**							
SOAPfuse	84	80	65	**88**						
Tophat	86	80	63	68	**87**					
ShortFuse	86	80	66	71	67	**87**				
MapSplice	76	73	62	60	64	60	**77**			
FusionCatcher	76	69	61	62	62	59	58	**76**		
FusionHunter	60	59	47	48	55	50	49	47	**61**	
FusionMap	29	26	27	20	19	20	24	21	13	**29**

MOJO Specificity



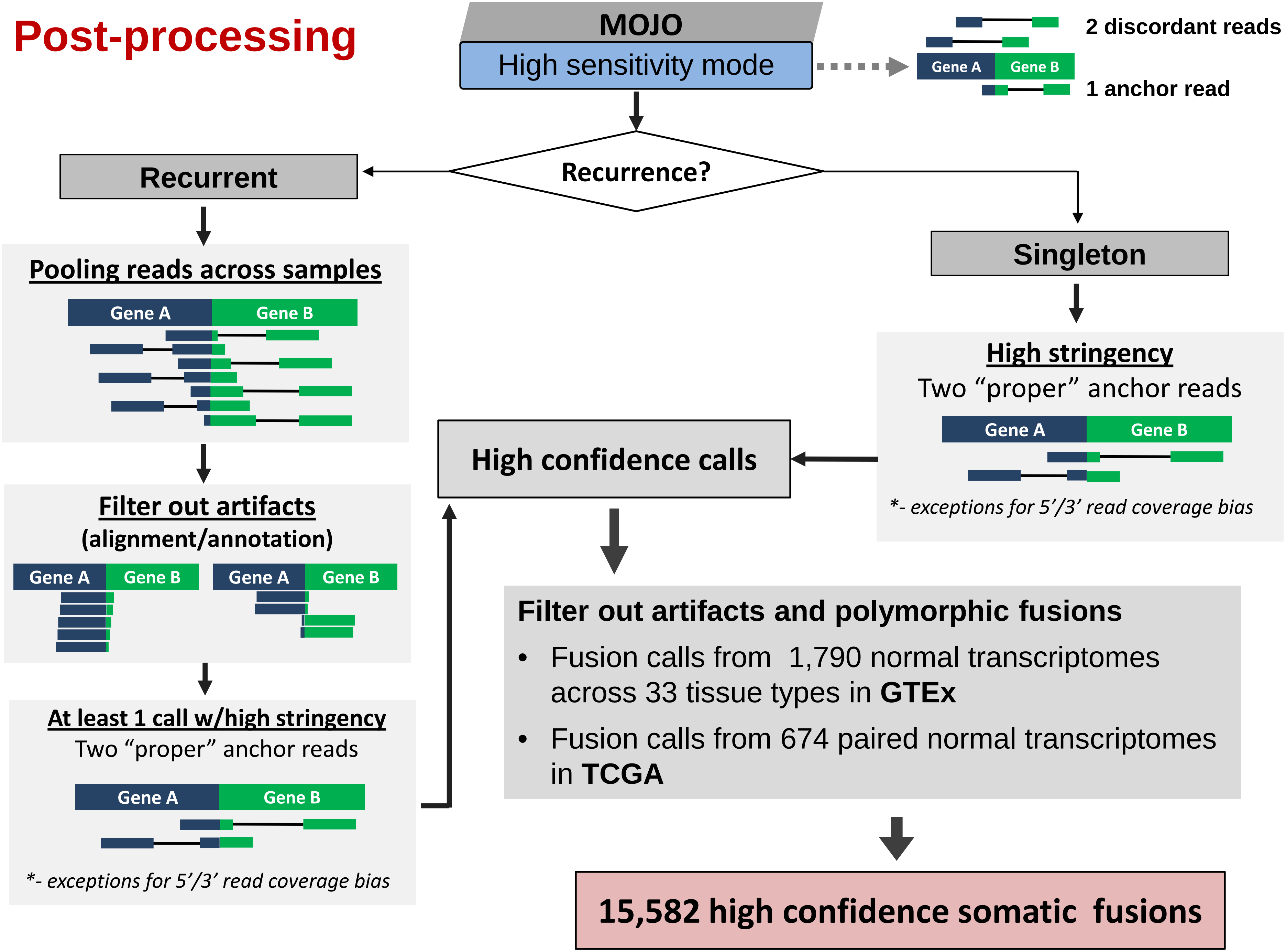


Fusion discovery across 7,175 tumor transcriptomes



Node Spec: 24 cores (@2.4Ghz), 32 GB RAM (“Beagle” – Cray XE6 supercomputer)

Post-processing



MOJO

High sensitivity mode

2 discordant reads

1 anchor read

Recurrence?

Recurrent

Pooling reads across samples

Filter out artifacts
(alignment/annotation)

At least 1 call w/high stringency
Two "proper" anchor reads

**- exceptions for 5'/3' read coverage bias*

Singleton

High stringency
Two "proper" anchor reads

**- exceptions for 5'/3' read coverage bias*

High confidence calls

Filter out artifacts and polymorphic fusions

- Fusion calls from 1,790 normal transcriptomes across 33 tissue types in **GTEx**
- Fusion calls from 674 paired normal transcriptomes in **TCGA**

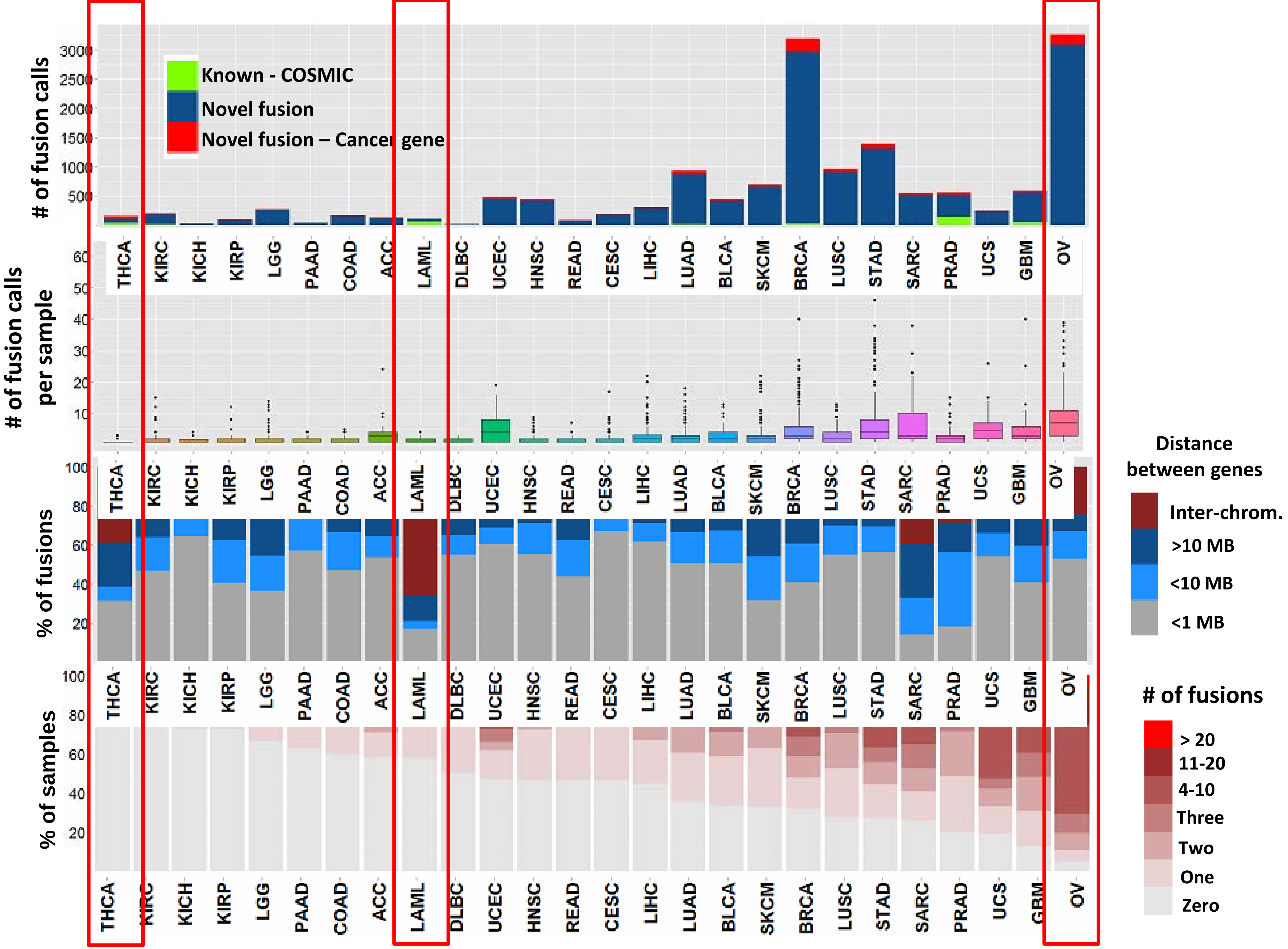
15,582 high confidence somatic fusions

Known recurrent fusions

	BLCA	BRCA	CESC	COAD	GBM	HNSC	KIRC	KIRP	LAML	LGG	LIHC	LUAD	LUSC	OV	PRAD	SARC	SKCM	STAD	THCA	UCEC	Total Calls	Evidence
Total	228	1041	173	264	169	425	513	170	172	369	166	513	483	416	280	105	374	266	505	156		
TMPRSS2_ERG															122						122	
PML_RARA									16												16	LAML marker paper
CBFB_MYH11									11												11	
RUNX1_RUNX1T1									7												7	
BCR_ABL1									4												4	
MLL_MLLT10									4												4	
MLL_ELL									3												3	
NUP98_NSD1									3												3	
MLL_MLLT3									2												2	
NSD1_NUP98									2												2	
PICALM_MLLT10									2												2	
MLL_MLLT3									1													
FGFR3_TACC3	7	1	1		34	2		2		2	1	1	5					2			58	GBM marker paper
EGFR_SEPT14					12	1				2											15	
LANCL2_SEPT14					7																7	
SEC61G_EGFR					5																5	
EGFR_SEC61G					3	1							1								5	
SFPQ_TFE3							5	1													6	KIRC marker
TRIP12_SLC16A14							2					1									3	
DHX33_NLRP1							2														2	
CCDC6_RET				1								1							22		24	
ETV6_NTRK3		1		2													1		6		10	
TPM3_NTRK1																1			2		3	
PAX8_PPARG																			5		5	
CD74_ROS1												2									2	
EML4_ALK												5							1		6	
TCF7L2_VTI1A											2			1				1		1	5	

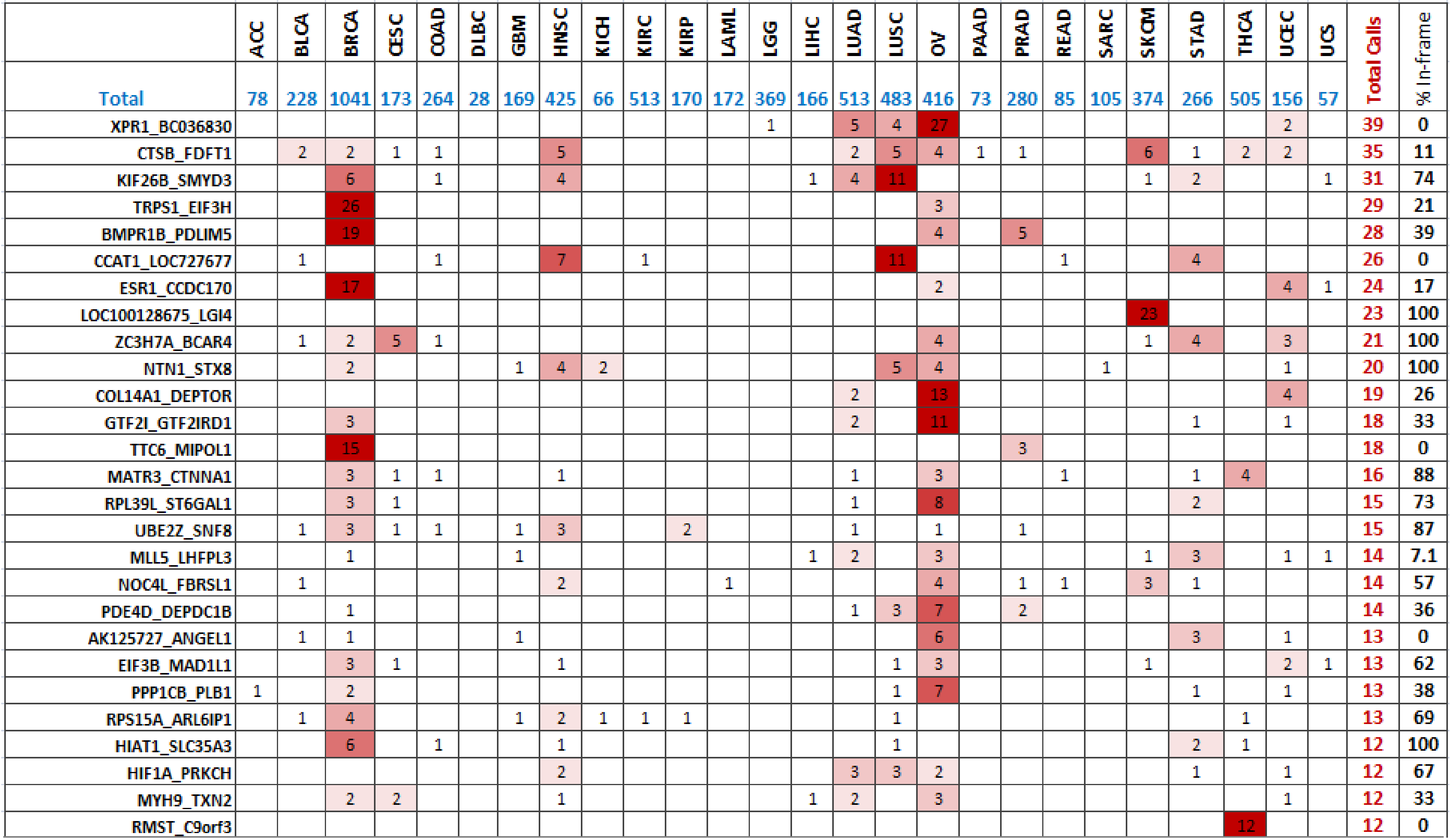
Non-canonical fusion

FGFR3::TACC3 in 20% of GBMs



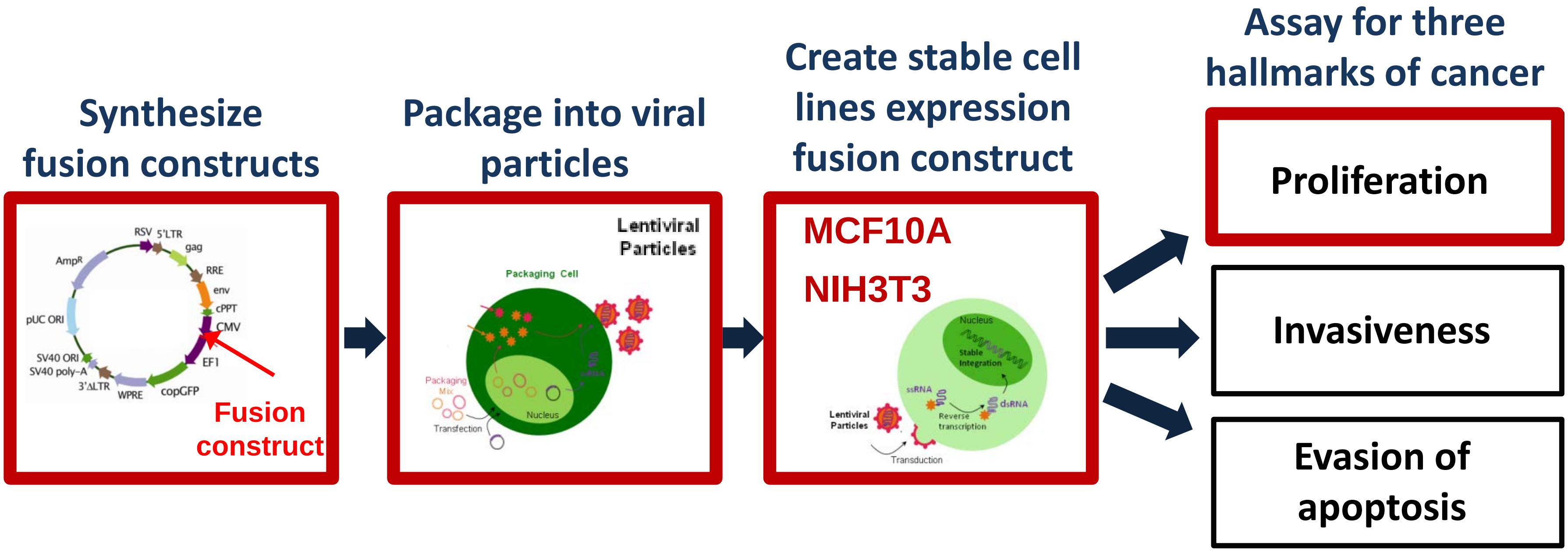
# of fusion calls
3000
2500
2000
1500
1000
500
Known - COSMIC
Novel fusion
Novel fusion – Cancer gene
# of fusion calls per sample
60
50
40
30
20
10
% of fusions
100
80
60
40
20
Distance between genes
Inter-chrom.
>10 MB
<10 MB
<1 MB
% of samples
100
80
60
40
20
# of fusions
> 20
11-20
4-10
Three
Two
One
Zero
THCA
KIRC
KICH
KIRP
LGG
PAAD
COAD
ACC
LAML
DLBC
UCEC
HNSC
READ
CESC
LIHC
LUAD
BLCA
SKCM
BRCA
LUSC
STAD
SARC
PRAD
UCS
GBM
OV

Novel recurrent fusions



	ACC	BLCA	BRCA	CESC	COAD	DLBC	GBM	HNSC	KICH	KIRC	KIRP	LAML	LGG	LIHC	LUAD	LUSC	OV	PAAD	PRAD	READ	SARC	SKCM	STAD	THCA	UCEC	UCS	Total Calls	% In-frame
Total	78	228	1041	173	264	28	169	425	66	513	170	172	369	166	513	483	416	73	280	85	105	374	266	505	156	57		
XPR1_BC036830													1		5	4	27								2		39	0
CTSB_FDFT1		2	2	1	1			5							2	5	4	1	1			6	1	2	2		35	11
KIF26B_SMYD3			6		1			4						1	4	11						1	2			1	31	74
TRPS1_EIF3H			26														3										29	21
BMPR1B_PDLIM5			19														4		5								28	39
CCAT1_LOC727677		1			1			7		1						11				1			4				26	0
ESR1_CCDC170			17														2								4	1	24	17
LOC100128675_LGI4																						23					23	100
ZC3H7A_BCAR4		1	2	5	1												4					1	4		3		21	100
NTN1_STX8			2				1	4	2							5	4				1				1		20	100
COL14A1_DEPTOR															2		13								4		19	26
GTF2I_GTF2IRD1			3												2		11						1		1		18	33
TTC6_MIPOL1			15																3								18	0
MATR3_CTNNA1			3	1	1			1							1		3			1			1	4			16	88
RPL39L_ST6GAL1			3	1											1		8						2				15	73
UBE2Z_SNF8		1	3	1	1		1	3			2				1		1		1								15	87
MLL5_LHFPL3			1				1							1	2		3					1	3		1	1	14	7.1
NOC4L_FBRSL1		1						2				1					4		1	1		3	1				14	57
PDE4D_DEPDC1B			1												1	3	7		2								14	36
AK125727_ANGEL1		1	1				1										6						3		1		13	0
EIF3B_MAD1L1			3	1				1								1	3					1			2	1	13	62
PPP1CB_PLB1	1		2													1	7						1		1		13	38
RPS15A_ARL6IP1		1	4				1	2	1	1	1					1								1			13	69
HIAT1_SLC35A3			6		1			1								1							2	1			12	100
HIF1A_PRKCH								2							3	3	2						1		1		12	67
MYH9_TXN2			2	2				1						1	2		3								1		12	33
RMST_C9orf3																								12			12	0

Functional validations workflow



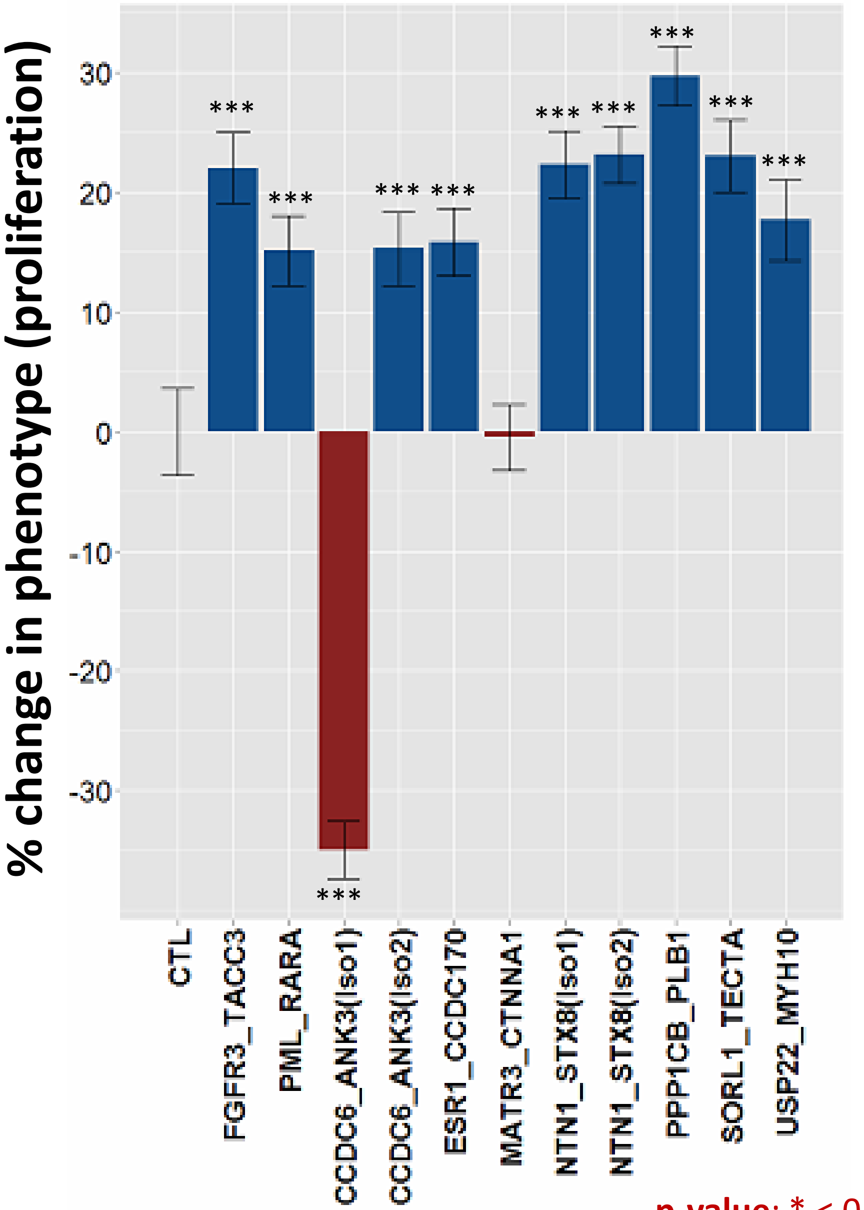
<u>Preliminary results:</u>

- 11 fusion constructs (9 fusion genes)
  - 7 novel fusions + 2 positive controls
- MCF10A
- Proliferation assays:
  - EdU
  - CellTiter Glo

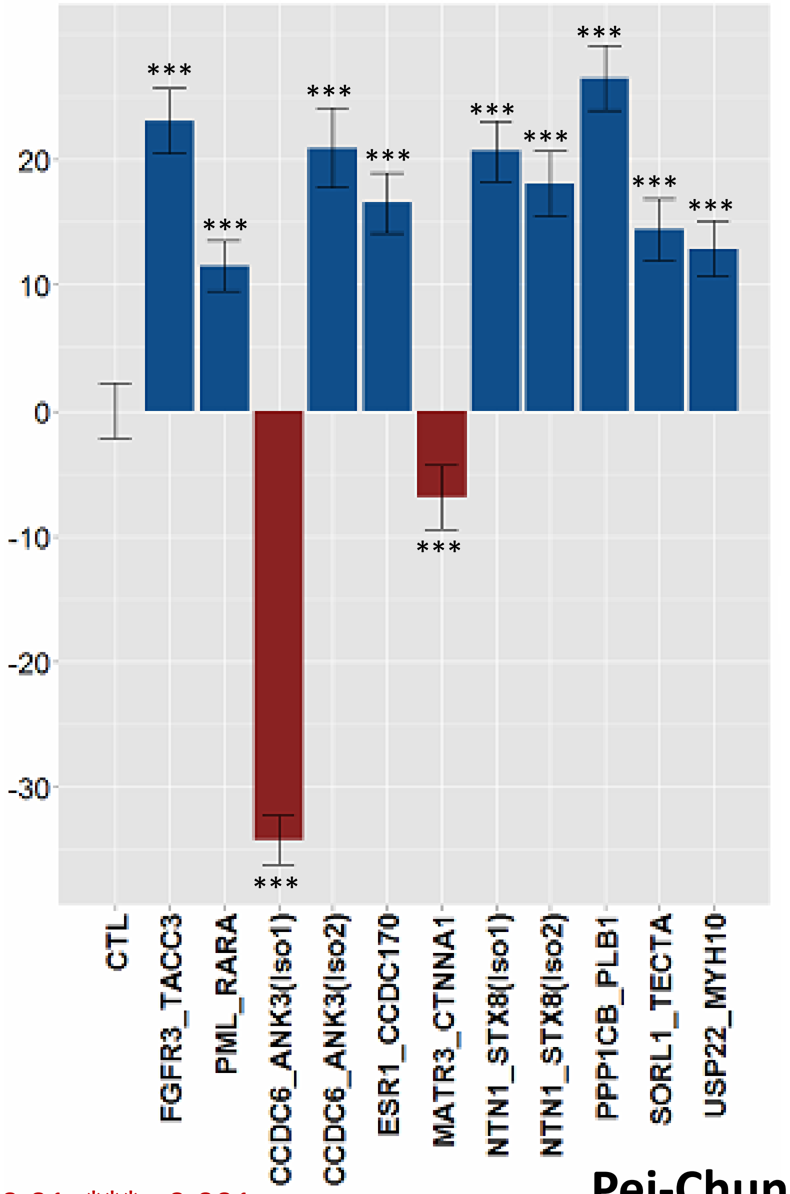
Pei-Chun Lin

Proliferation assays MCF10A

EdU



CellTiter Glo



p-value: * < 0.05, ** < 0.01, *** <0.001

Pei-Chun Lin

Future directions

- Functional validations
  - Using a different cell line (NIH3T3)
  - Over-expressing full length individual genes as controls
  - Further characterization of functional fusions
- Recurrently fused genes are of significant interest
- Integrated analysis to find associations between copy number alterations, mutations and fusion events

Acknowledgements

Pei-Chun Lin

Kevin White

Jason Pitt
Michael Bolt
Megan McNerney
Jason Grundstad
Audrey Fu

Michelle LeBeau

Funding:

- NIH Genetics & Regulation Training Grant

Institute for
Genomics &
Systems Biology

Bionimbus
Allison Heath
Robert Grossman

“Beagle” – Argonne National Labs
Lorenzo Pesce
Joe Urbanski

Recombineering
Jennifer Moran (UChicago)
Jingdong Tian (Duke)

Cellular Screening Facility
Siquan Chen
Sam Bettis

BC Genome Sciences Center
Andy Mungall
Reanne Bowlby
Karen Mungall