ICGC Pan-cancer analysis of whole genomes:
Integration of transcriptome and genome working group (PCAWG-3)

Angela Brooks
Postdoctoral Fellow
Lab of Matthew Meyerson

GDC Call

December 18, 2014

* Group Leaders

PCAWG-3 Group

Broad Institute
Matthew Meyerson*
Angela Brooks*
Gaddy Getz
Chandra Pedamallu
Sam Freeman
David DeLuca
Ayellet Segre
Tim Sullivan
Lihua Zou

EMBL-EBI
Alvis Brazma*
Nuno Fonseca
Lilian Greger
Mar Gonzàlez-Porta
Oliver Stegle

MSKCC
Gunnar Rätsch*
Andre Kahles
Kjong Lehmann
Natalie Davidson
Stefan Stark

University of Chicago
Zhenyu Zhang
Allison Heath
Bob Grossman

UC Santa Cruz
Kyle Ellrott
Christopher Wilks

University of Tokyo
Yuichi Shiraishi

Ontario Institute for Cancer Research
Francis Ouellette
Marc Perry

Baylor College of Medicine
Chad Creighton
Yuan Yuan

UNC Chapel Hill
Matthew Wilkerson
Katherine Hoadley

Hospital for Sick Children
Adam Shlien

MD Anderson
John Zhang
Ken Chen
Leng Han
Sahil Seth
Wanding Zhou
Xian Fan
Zechen Chong
Samir Amin
Han Liang

EMBL Heidelberg
Alejandro Reyes
Jan Korbel

Yale
Mark Gerstein
Anurag Sethi

Washington University in St. Louis
Reyka Jayasinghe
Venkata Yellapantula

Dana-Farber
Virginia Savova

Genome Institute of Singapore
Jonathan Goke
Tannistha Nandi
Patrick Tan

Sage Bionetworks
Yin Hu

NorthShore University Health System
Yuan Ji

BGI
Yong Hou

Weill Cornell Medical College
Ekta Khurana

Mission of PCAWG-3: Integration of transcriptome and genome

- Scientific mission
  - Characterizing genomic alterations that lead to transcriptome alterations and contribute to cancer phenotypes
    - Somatic and germline variants

- Data deliverables
  - Provide a unified analysis of PCAWG RNA-Seq data
    - 1,000+ RNA-Seq samples
  - Identify and quantify transcriptome-level cancer genome alterations

Transcriptome group deliverables

- Alignment of PCAWG RNA-Seq
- RNA-Seq alignments/coverage with no sequence information
- Exon expression
- Splice junction expression
- Transcript expression
- Gene expression
- Splicing event quantification
- RNA-Seq variant calls
- RNA fusion

DCC RNA-Seq Samples in Release 17

Source	Normal	Tumor	Total
ICGC	0	120	120
TCGA	64	841	905
Total	**64**	**961**	**1025**

Marc Perry, Francis Ouellette, DCC/OICR
Chad Creighton, Baylor College of Medicine

Challenges in building the RNA-Seq alignment SOP

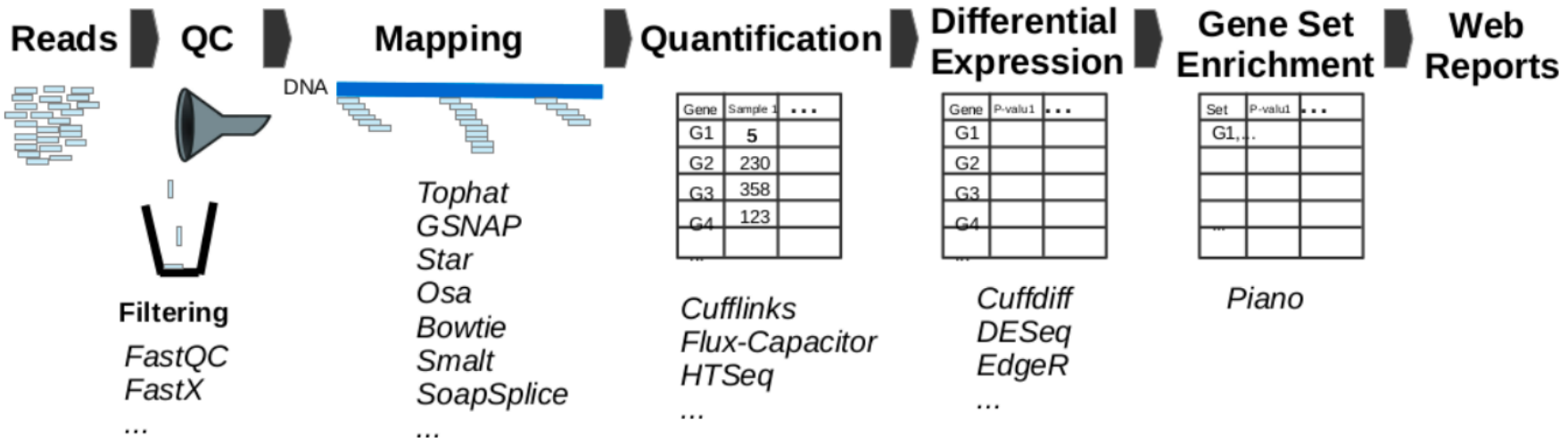
1. No clear consensus of “best” aligner tool
2. Data is not standardized: TCGA and ICGC RNA-Seq are contributed by different groups
   - e.g., data provided as aligned BAM or FASTQ, mixed sequencing platforms
3. Sorting out computational resources and data distribution
4. Curating and associating meta-data to alignment files

RNA-Seq analysis assessments from members of PCAWG-3

- **EMBL:** Nuno Fonseca, John Marioni, Alvis Brazma
  - Presentation:
    - https://wiki.oicr.on.ca/download/attachments/58396231/pipeline_comparison_PAWG_Fonseca_Brazma.pdf?version=1&modificationDate=1403813779000&api=v2
- **MSKCC:** Andre Kahles, Kjong Lehmann, Gunnar Rätsch
  - Presentation:
    - https://docs.google.com/presentation/d/1P9Z0LbprPQ1n0rAlS_1bJlRClQ--H54kHjopvuiXk98/edit?usp=sharing
- **Broad Institute/GTEx:** David DeLuca, Tim Sullivan
  - Presentation:
    - https://wiki.oicr.on.ca/download/attachments/58396231/GTEx_Aligner_Benchmarking_06-26-14.pdf?version=1&modificationDate=1403813779000&api=v2

RNA-Seq analysis assessments from members of PCAWG-3

iRAP- an integrated RNA-Seq analysis pipeline
http://biorxiv.org/content/early/2014/06/06/005991.full-text.pdf+html



EMBL
Nuno Fonseca, John Marioni, Alvis Brazma

RNA-Seq analysis assessments from members of PCAWG-3

Ranking of pipelines for gene expression quantification based on simulated data

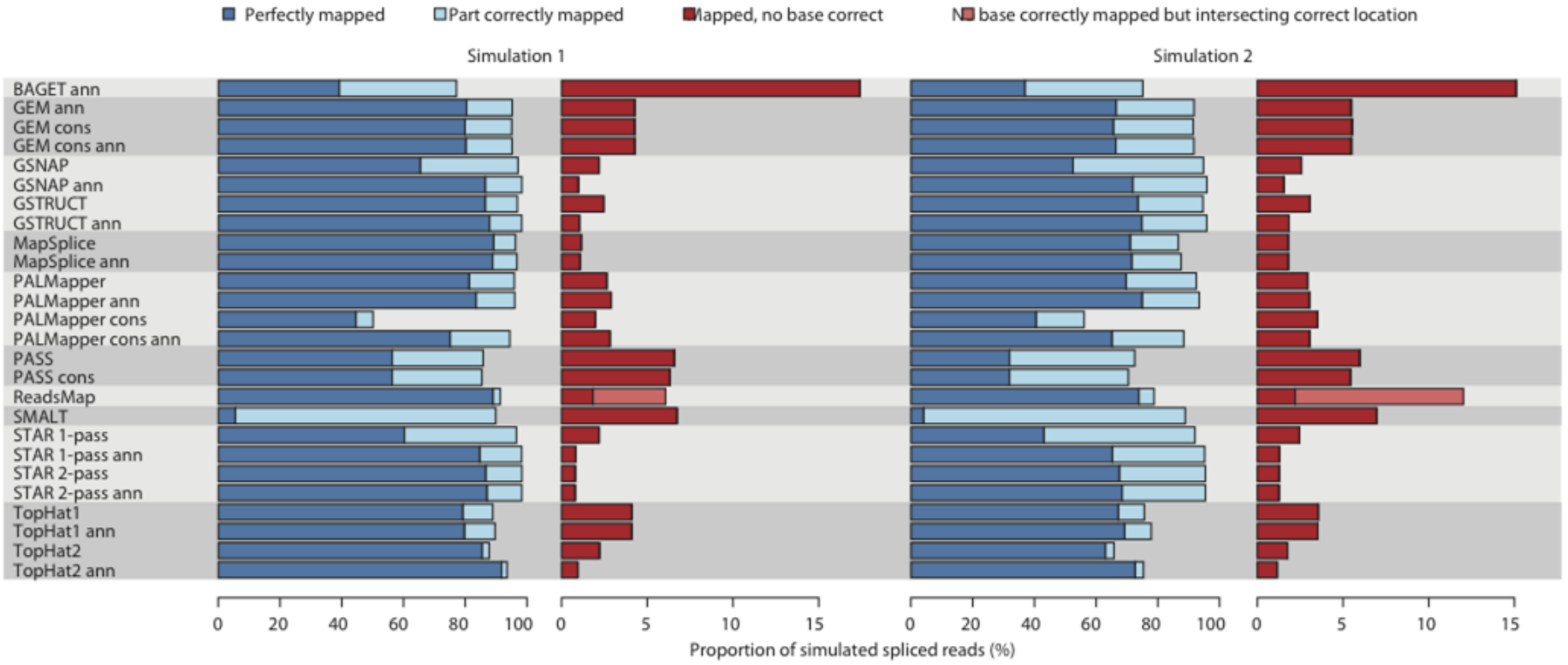
Pipeline		Overall	Error		Spearman	
Aligner	Quant. Method	Rank	Avg. Rank	mean ± sd	Avg. Rank	mean ± sd
osa	htseq-ine	17	9	16.88 ± 2.00	8	0.94 ± 0.01
tophat1	htseq-ine	18	10	17.28 ± 2.52	8	0.94 ± 0.01
gsnap	htseq-ine	23	11	19.87 ± 7.78	12	0.93 ± 0.01
tophat2	htseq-ine	23	13	19.10 ± 5.93	10	0.94 ± 0.01
osa	fluxcapacitor	24	22	19.95 ± 2.04	2	0.95 ± 0.00
star	htseq-ine	24	11	17.37 ± 2.81	13	0.93 ± 0.01
smalt	htseq-ine	25	14	18.93 ± 7.22	11	0.93 ± 0.00
tophat1	fluxcapacitor	27	23	19.79 ± 2.08	4	0.94 ± 0.00
bwa2	htseq-ine	28	16	20.34 ± 6.05	12	0.93 ± 0.02
star	fluxcapacitor	29	23	20.51 ± 2.72	6	0.94 ± 0.00
tophat1	cufflinks2	31	13	22.74 ± 16.44	18	0.92 ± 0.03
osa	cufflinks2	33	13	20.87 ± 9.57	20	0.92 ± 0.03
star	cufflinks2	34	12	16.94 ± 2.96	22	0.92 ± 0.01
tophat2	fluxcapacitor	34	26	22.17 ± 4.44	9	0.94 ± 0.01
bwa2	fluxcapacitor	36	26	20.84 ± 2.99	10	0.94 ± 0.01
osa	htseq-u	36	16	22.07 ± 15.13	21	0.92 ± 0.02
tophat1	cufflinks1	36	13	22.35 ± 15.58	23	0.92 ± 0.01
tophat1	htseq-u	36	17	19.84 ± 4.83	19	0.92 ± 0.00
bwa2	htseq-u	37	20	23.75 ± 13.71	17	0.92 ± 0.02
gsnap	cufflinks2	37	16	24.26 ± 17.44	22	0.91 ± 0.05
gsnap	fluxcapacitor	37	26	22.71 ± 6.43	11	0.94 ± 0.00
smalt	htseq-u	37	19	20.30 ± 7.67	19	0.92 ± 0.01
osa	cufflinks1	38	13	21.12 ± 12.17	25	0.92 ± 0.01
bwa1	htseq-ine	39	21	23.84 ± 8.04	18	0.92 ± 0.03
star	cufflinks1	39	12	17.65 ± 4.28	28	0.91 ± 0.01
tophat2	cufflinks2	39	14	23.35 ± 16.97	25	0.91 ± 0.07
gsnap	htseq-u	40	16	23.46 ± 13.44	24	0.92 ± 0.01
tophat2	htseq-u	40	19	21.44 ± 7.53	21	0.92 ± 0.01

- Average rank based on 32 data sets of differing read length, sequencing depth, and single/paired-end reads
- Overall rank = error rank + spearman rank

EMBL
Nuno Fonseca, John Marioni, Alvis Brazma

RNA-Seq analysis assessments from members of PCAWG-3

RGASP3 (RNA-seq Genome Annotation Assessment Project):
Read placement accuracy for simulated spliced reads

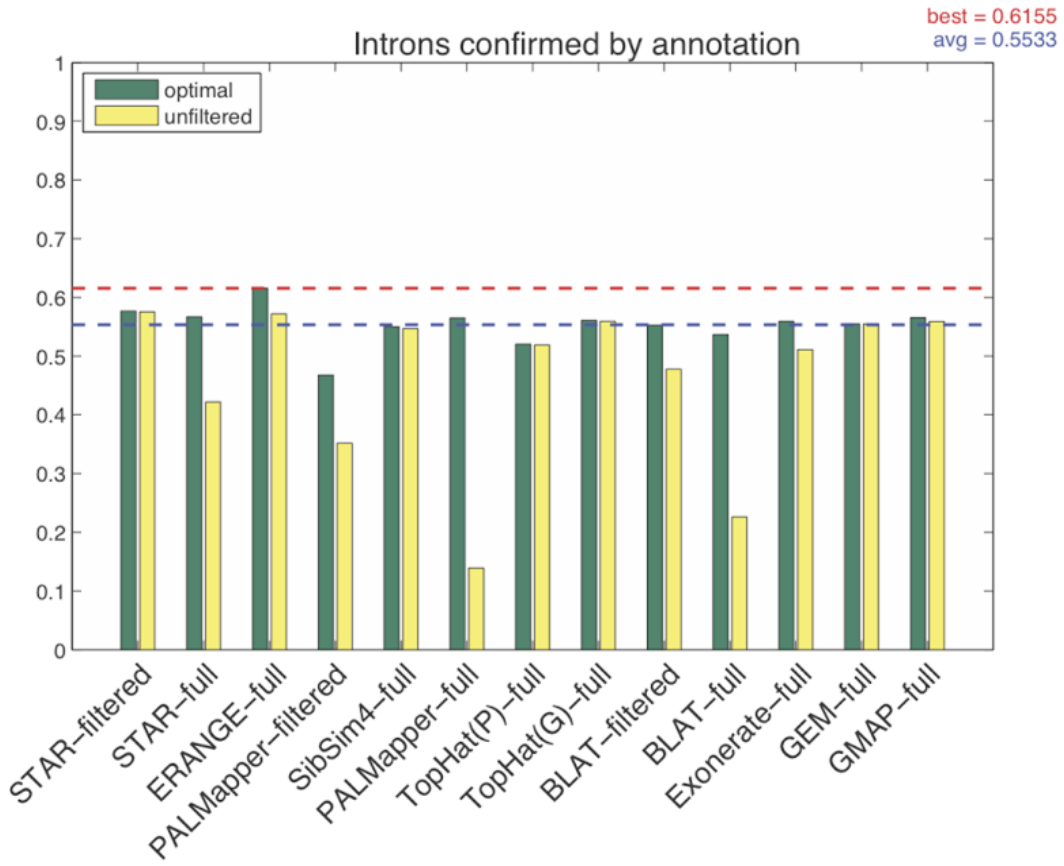


Engstrom et al., Nature Methods 2013

MSKCC
Andre Kahles, Kjong Lehmann, Gunnar Rätsch

RNA-Seq analysis assessments from members of PCAWG-3

Intron accuracy is improved by filtering



Filters:

- Number of edit-operations
- Minimum segment length
- Number of spliced reads

MSKCC

Andre Kahles, Kjong Lehmann, Gunnar Rätsch

RNA-Seq analysis assessments from members of PCAWG-3

Many more analysis details…

Systematic evaluation of spliced alignment programs for RNA-seq data

Pär G Engström[1,13], Tamara Steijger[1], Botond Sipos[1], Gregory R Grant[2,3], André Kahles[4,5], The RGASP Consortium[6], Gunnar Rätsch[4,5], Nick Goldman[1], Tim J Hubbard[7], Jennifer Harrow[7], Roderic Guigó[8,9] & Paul Bertone[1,10–12]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge, UK. [2]Department of Genetics, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [3]Penn Center for Bioinformatics, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [4]Computational Biology Center, Sloan-Kettering Institute, New York, New York, USA. [5]Friedrich Miescher Laboratory of the Max Planck Society, Tübingen, Germany. [6]Full lists of members and affiliations appear at the end of the paper. [7]Wellcome Trust Sanger Institute, Cambridge, UK. [8]Centre for Genomic Regulation, Barcelona, Spain. [9]Universitat Pompeu Fabra, Barcelona, Spain. [10]Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany. [11]Developmental Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany. [12]Wellcome Trust–Medical Research Council Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK. [13]Present address: Department of Biochemistry and Biophysics, Science for Life Laboratory, Stockholm University, Stockholm, Sweden. Correspondence should be addressed to P.B. (bertone@ebi.ac.uk).

RECEIVED 31 MARCH; ACCEPTED 10 SEPTEMBER; PUBLISHED ONLINE 3 NOVEMBER 2013; DOI:10.1038/NMETH.2722

NATURE METHODS | ADVANCE ONLINE PUBLICATION | 1

MSKCC

Andre Kahles, Kjong Lehmann, Gunnar Rätsch

RNA-Seq analysis assessments from members of PCAWG-3

- GTEx objectives
  - Gene expression
  - eQTL identification
  - Isoform reconstruction and quantification
  - Allele specific expression (ASE)
  - Production in a scalable manner

Broad Institute/GTEx
Tim Sullivan, David DeLuca

RNA-Seq analysis assessments from members of PCAWG-3

- Aligner comparisons based on simulated data
  - TopHat2
    - Best for split reads (intron spanning reads), especially when read spans two introns
  - GEM
    - Best for accurate read mapping in the presence of polymorphism/sequencing errors
  - STAR
    - Amazing fast, not good for novel junctions, decent for other metrics
  - STAR 2 pass
    - Almost as fast, recovers loss of performance in novel isoform species

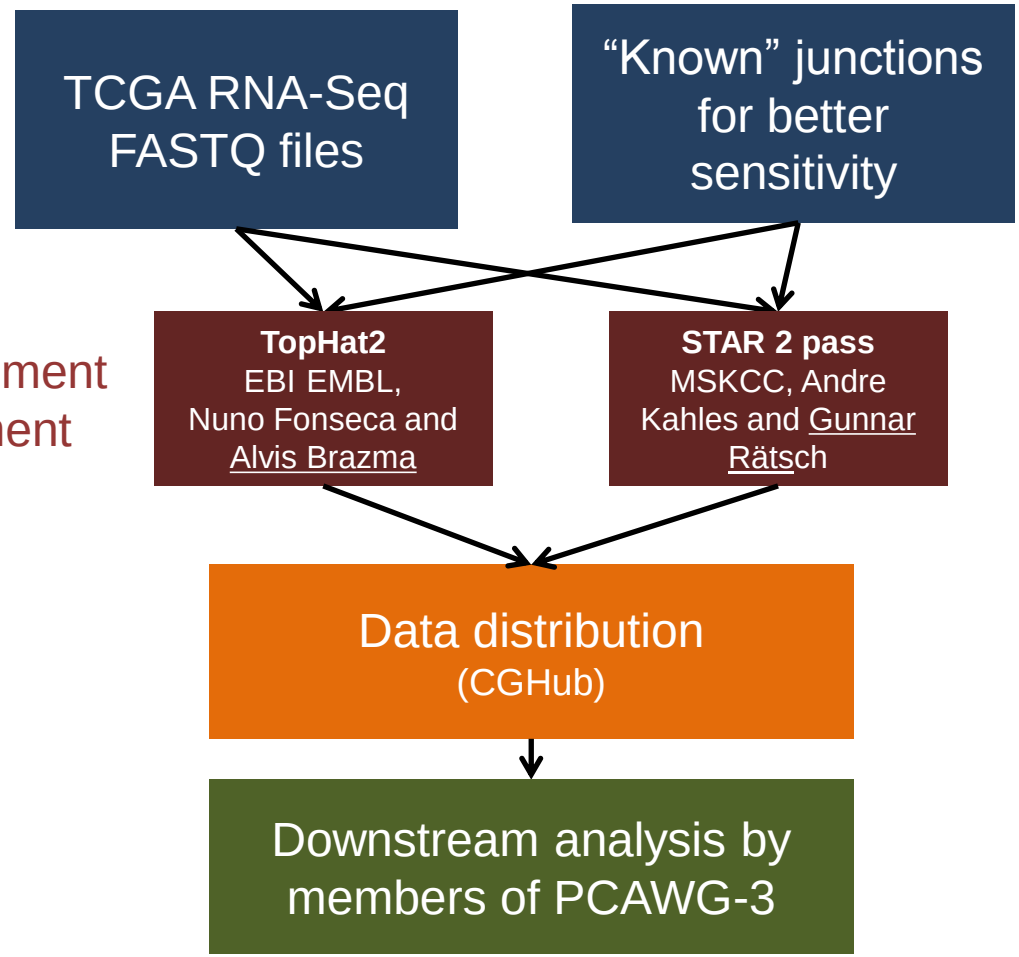
Broad Institute/GTEx
Tim Sullivan, David DeLuca

Conclusion from analysis assessments

- No aligner gives the best results for all metrics
- Choice of aligner and downstream analysis tools depend on what is most important for specific deliverable
  - Prioritized deliverables
    - Exon level quantification
    - Gene expression
    - Splice junction detection

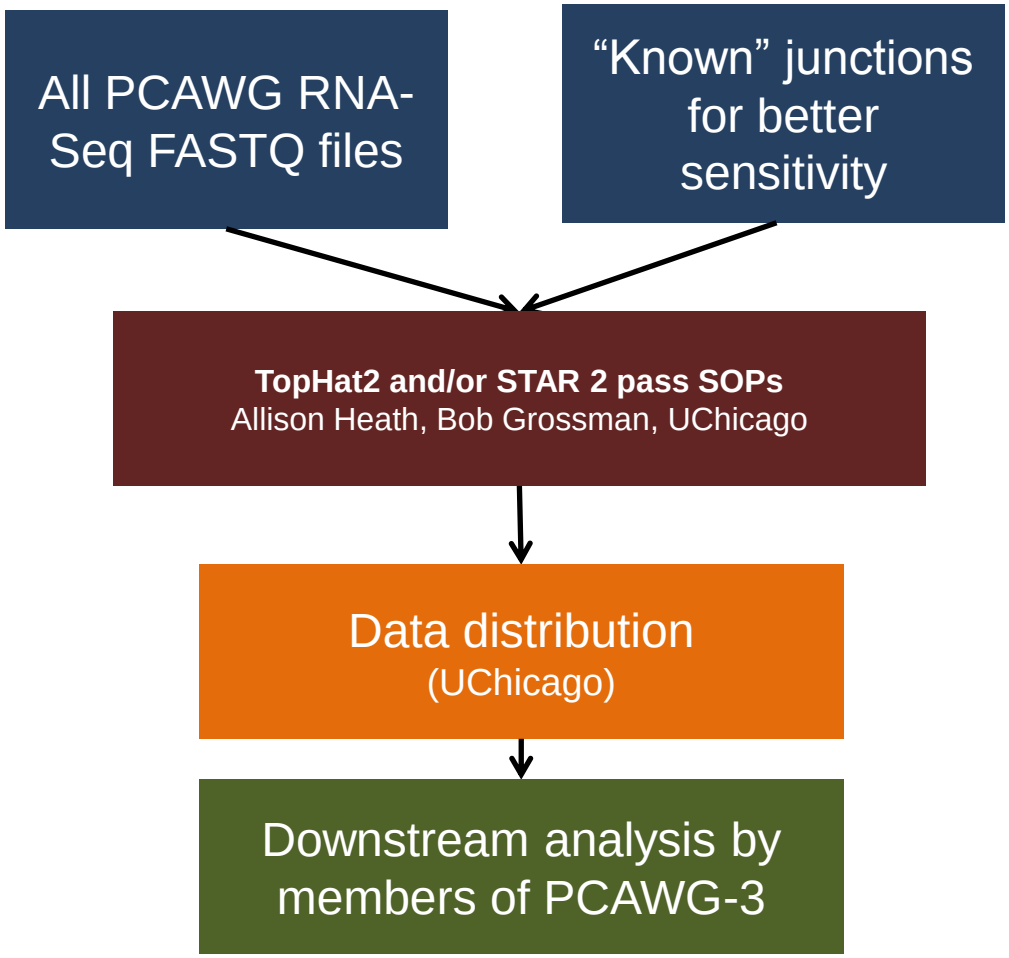
Current RNA-Seq SOP development plans for PCAWG-3

RNA-Seq alignment SOP development



- Potentially use both aligners on all data or select one based on downstream analysis
  - ENCODE is producing alignments from TopHat2 and STAR

Future RNA-Seq SOP plan for PCAWG-3



Downstream analysis tools

Analysis	Tool/Method	Group
Gene quantification	Htseq-count	Brazma
	RSEM	Gerstein/ENCODE pipeline
Exon quantification	bedtools	Stuart
Junction quantification	Custom	Stuart
Transcript structure	RNA Architect	Shlien
RNA editing	Custom tool	Hery Yang, Jun Wang
Allele-specific expression	UNCeqR	Wilkerson
Alternative promoters	Custom R/Bioconductor	Tan
lncRNA	Cufflinks, Cuffdiff	Brazma
	Cufflinks, HTSeq	Samir Amin, Chin
Transcript structure	RNA Architect	Shlien
	MiTie	Rätsch
Alternative splicing	JuncBASE	Meyerson
	DEXSeq, SwitchSeq	Brazma
	Cufflinks, Diffsplice	Guinney, Sage Bionetworks
	SplAdder/rDiff, Limix (sQTL)	Rätsch
Fusion transcripts	FusionSeq	Gerstein
	deFuse, Fusion Map	Brazma
	BreakTrans	Ken Chen