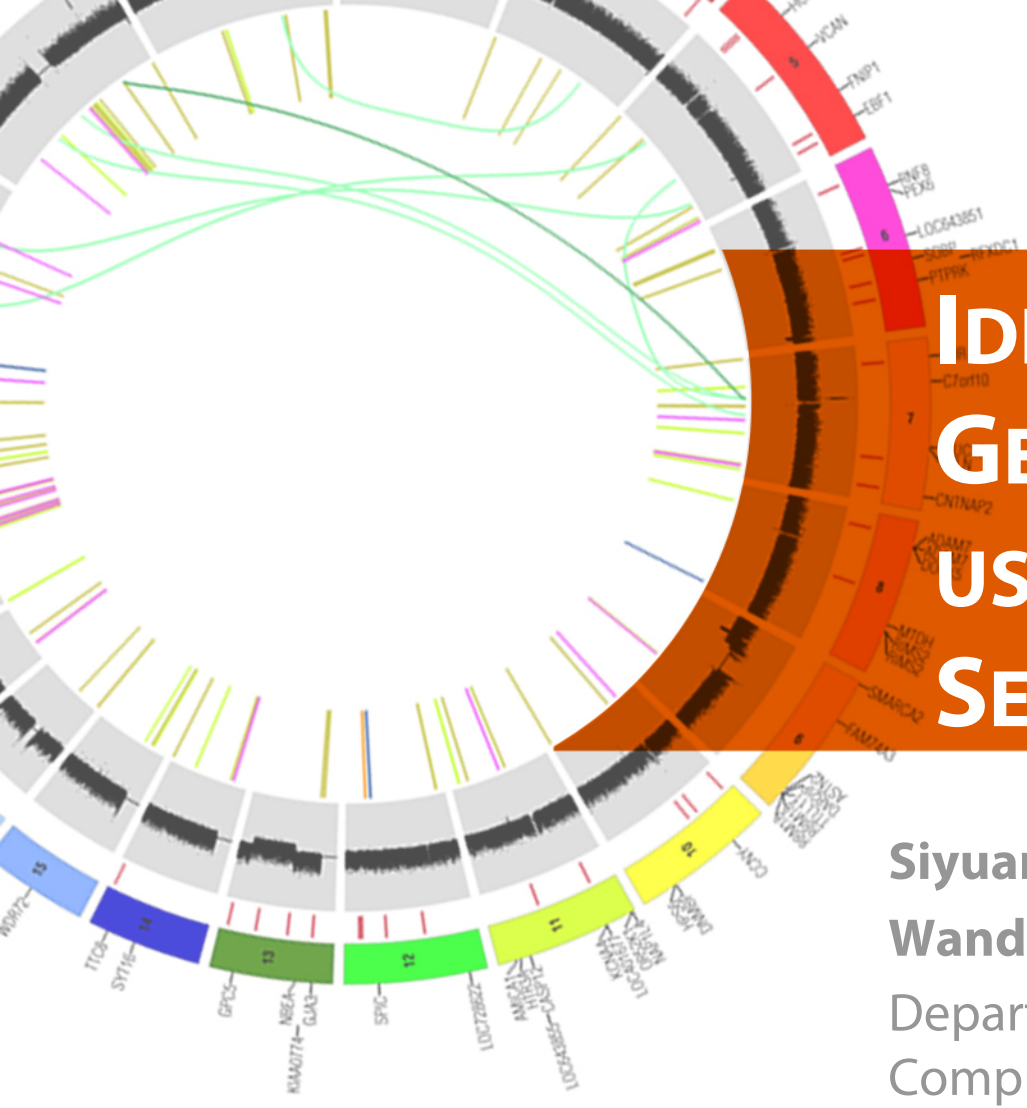


The Cancer Genome Atlas

Identification of Gene Fusions using RNA Sequencing Data

Siyuan Zheng

Wandaliz Torres-Garcia

Department of Bioinformatics and Computational Biology, MD Anderson Cancer Center

Roel Verhaak Lab

Nov 28, 2012

Poster #107

Gene fusion has been recognized in cancer as driver and target

A cellular oncogene is translocated to the Philadelphia chromosome in chronic myelocytic leukaemia

Annelies de Klein*, Ad Geurts van Kessel*, Gerard Grosveld*, Claus R. Bartram*, Anne Hagemeijer*, Dirk Bootsma*, Nigel K. Spurr†, John Groffen‡ &

RESEARCH ARTICLE

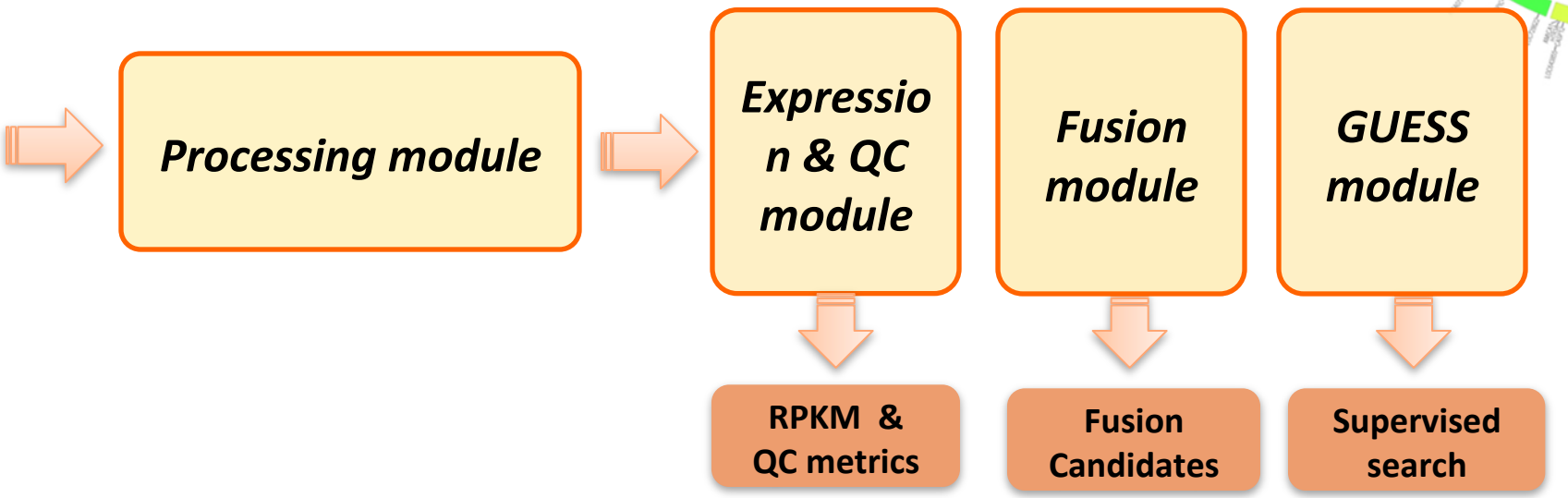
Recurrent Fusion of *TMPRSS2* and ETS Transcription Factor Genes in Prostate Cancer

Scott A. Tomlins,[1] Daniel R. Rhodes,[1,2] Sven Perner,[7,9] Saravana M. Dhanasekaran,[1] Rohit Mehra,[1] Xiao-Wei Sun,[7] Sooryanarayana Varambally,[1,6] Xuhong Cao,[1] Joelle Tchinda,[7] Rainer Kuefer,[10] Charles Lee,[7] James E. Montie,[3,5,6] Rajal B. Shah,[1,3,5,6] Kenneth J. Pienta,[3,4,5,6] Mark A. Rubin,[7,8] Arul M. Chinnaiyan[1,2,3,5,6*]

(*6*). This karyotypic complexity is thought to reflect secondary genomic alterations acquired during tumor progression.

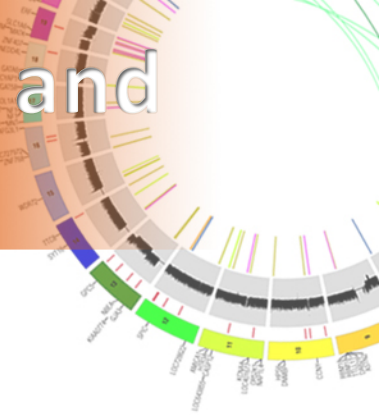
We hypothesized that rearrangements and high-level copy number changes that result in marked overexpression of an oncogene should be evident in DNA microarray data but not necessarily by traditional analytical approaches. In the majority of cancer types, heterogeneous patterns of oncogene activation have been observed; thus, traditional analytical methods that search for common activation of genes across a class of cancer samples (e.g., *t* test or signal-to-noise ratio) will fail to find such oncogene expression profiles. Instead, a meth-

PRADA: *P*ipeline for *R*N*A*-seq *D*ata *A*nalysis



- Four modules for paired-end RNA-sequencing data:
  1. Processing (read alignment, recalibration etc.)
  2. Expression calculation and Quality Control
  3. Gene fusion identification
  4. GUESS: ***G***eneral ***U***ser d***E***fined ***S***upervised ***S***earch

PRADA aligns reads to transcriptome and genome



- A multi-tiered alignment strategy
  - **Step I**: Map reads to reference comprised of transcriptome and genome.
  - **Step II**: Map transcript placements to genomic coordinates.
  - **Step III**: Filter reads with ambiguous placement on genome and isolate read pairs for future use.
- Advantages
  - Mapping to transcriptome captures all transcript variants.
  - Mapping to genome captures unannotated transcripts.

For more details, see Berger M. et al. Genome Res, 2010

PRADA leverages established tools as infrastructure

BAM

FASTQ

INPUTS

Config: Reference data

Processing Module

Read Alignment → Remap alignments → Combine two ends → Score recalibration

Expression & QC Module

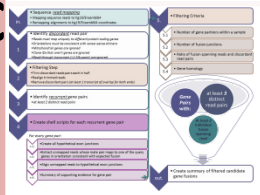
[YES|| NO|| ONLY]

RNA-SeQC

BROAD INSTITUTE

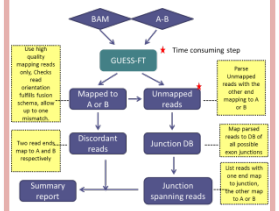
Fusion Module

[YES|| NO|| ONLY]



GUESS

[YES|| NO|| ONLY]



BWA | Samtools | Picard | GATK | RNA-SeQC

RPKM values were used to call GBM subtypes

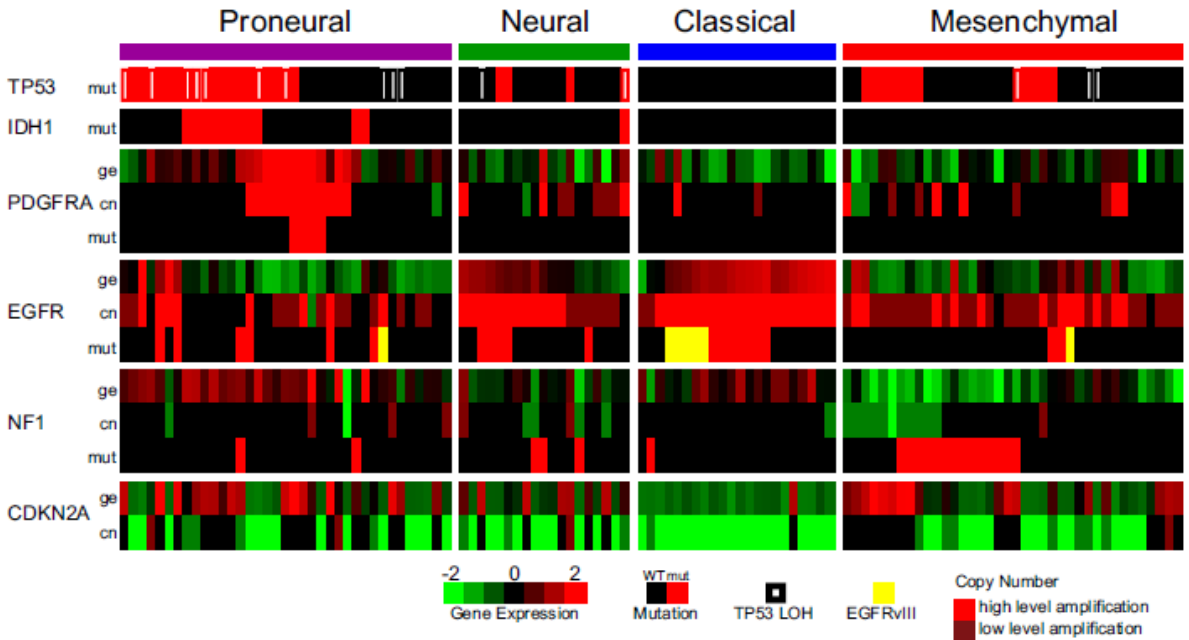
Rea
gen

U133A
based
calls

Cancer Cell
Article

Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in *PDGFRA*, *IDH1*, *EGFR*, and *NF1*

Roel G.W. Verhaak,[1,2,17] Katherine A. Hoadley,[3,4,17] Elizabeth Purdom,[7] Victoria Wang,[8] Yuan Qi,[4,5] Matthew D. Wilkerson,[4,5] C. Ryan Miller,[4,6] Li Ding,[9] Todd Golub,[1,10] Jill P. Mesirov,[1] Gabriele Alexe,[1] Michael Lawrence,[1,2] Michael O'Kelly,[1,2] Pablo Tamayo,[1] Barbara A. Weir,[1,2] Stacey Gabriel,[1] Wendy Winckler,[1,2] Supriya Gupta,[1] Lakshmi Jakkula,[11] Heidi S. Feiler,[11] J. Graeme Hodgson,[12] C. David James,[12] Jann N. Sarkaria,[13] Cameron Brennan,[14] Ari Kahn,[15] Paul T. Spellman,[11] Richard K. Wilson,[9] Terence P. Speed,[7,16] Joe W. Gray,[11] Matthew Meyerson,[1,2] Gad Getz,[1] Charles M. Perou,[3,4,8] D. Neil Hayes,[4,5,*] and The Cancer Genome Atlas Research Network



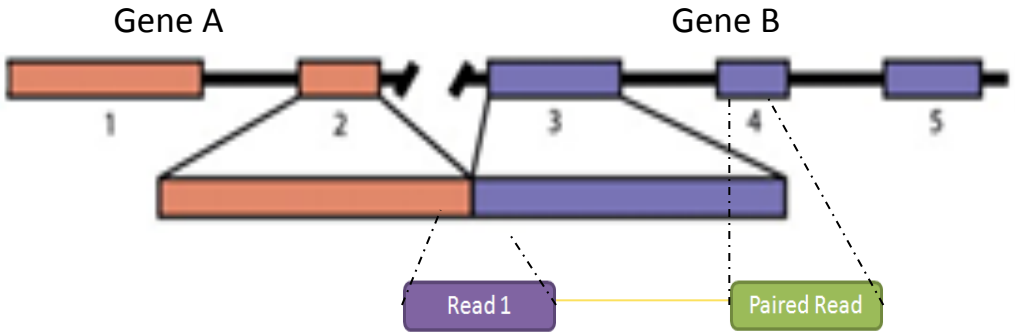
mal

nome Atlas

Rationale for fusion detection in PRADA

Discordant read pair

Fusion spanning reads



Fusion candidates

Figure adapted from Berger M. et al. Genome Res, 2010

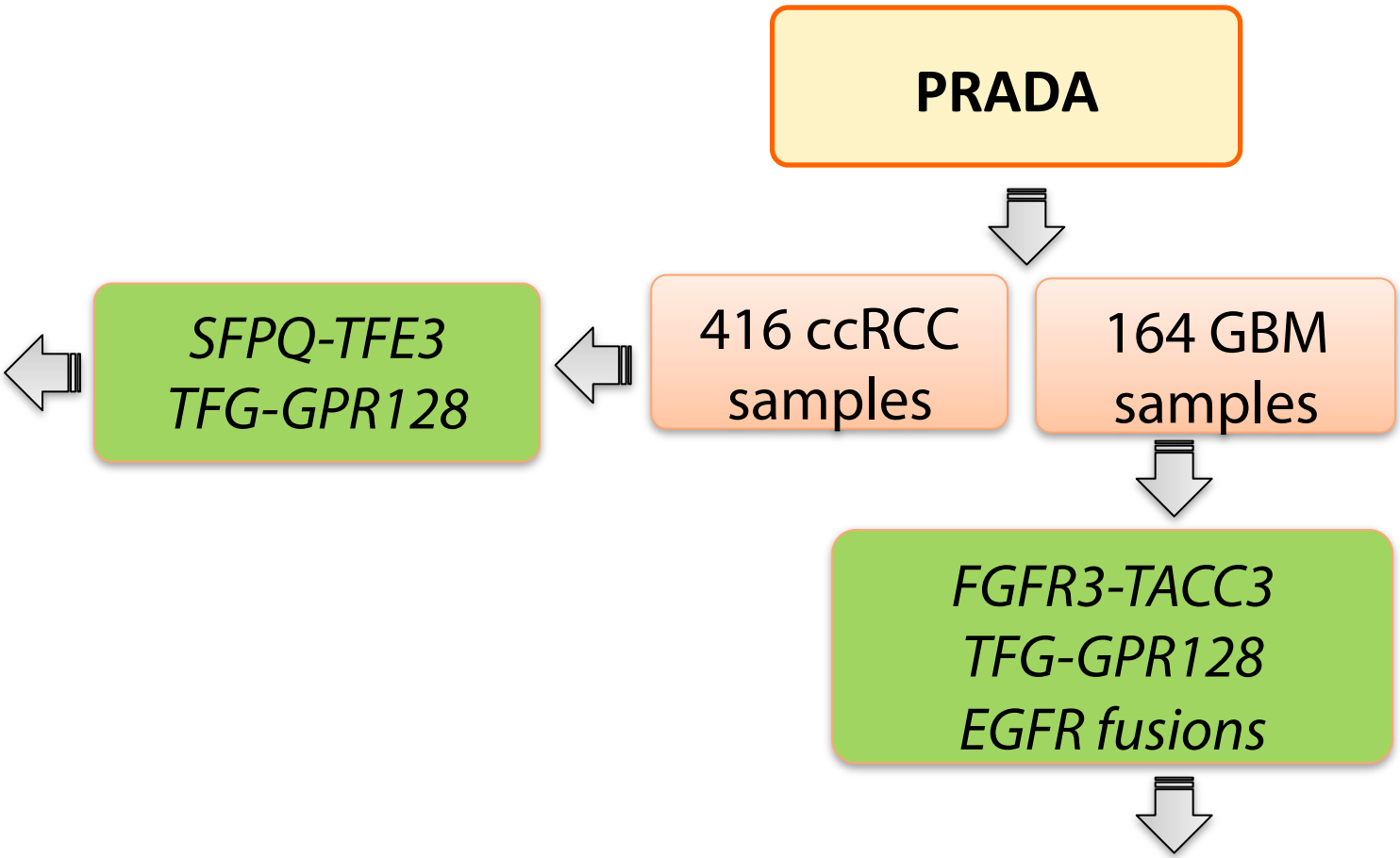
Filters are designed to exclude false positives

I. Homology filter
   - Gene partners should not have significant sequence similarity.

II. Ratio of fusion spanning reads over discordant reads
   - The ratio should be in a range, which is determined by library size and read length.

III. Minor filters that look at gene partners, junction pattern, etc.

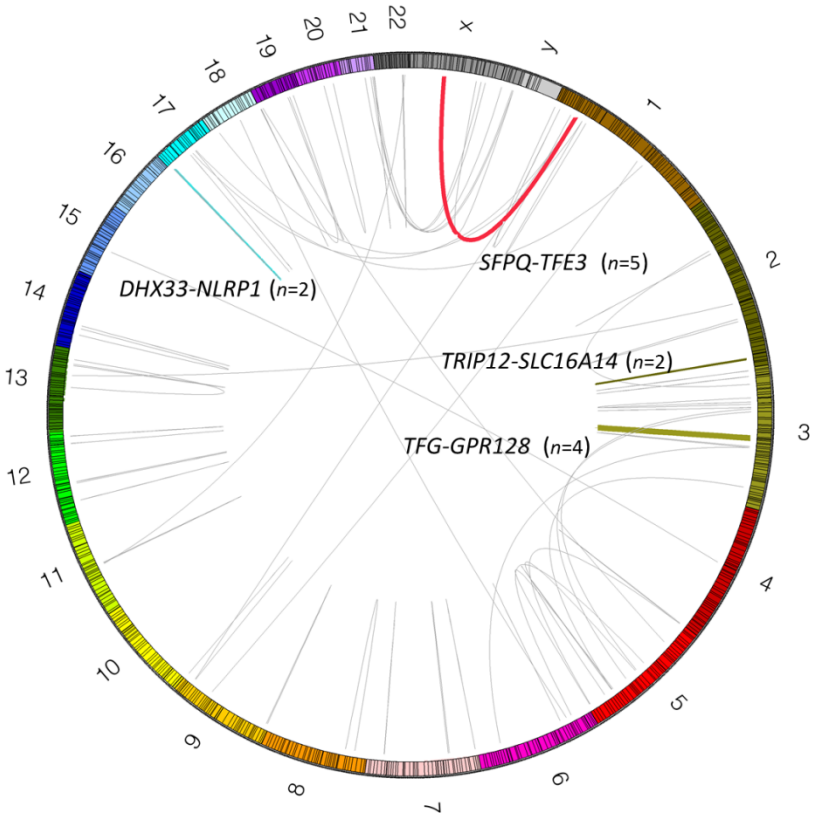
PRADA was applied to ccRCC and GBM



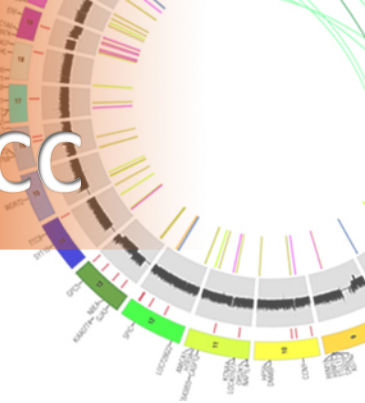
ccRCC: clear cell renal cell carcinoma (KIRC in TCGA)

Gene fusion results in ccRCC

- 80 fusions were found in 416 ccRCCs, 14.9% of the samples have at least 1 fusion.
- Recurrent fusions
  - *SFPQ-TFE3* (*n*=5, chr1-chrX)
  - *TFG-GRP128* (*n*=4, chr3)
  - *DHX33-NLRP1* (*n*=2, chr2)
  - *TRIP12-SLC16A14* (*n*=2, chr17)
- *TFE3* translocation is related to a rare subset of adult kidney cancer.
  Argani P. et al. 2004



Fusion validation using RT-PCR in ccRCC



Validation rate ≥ 85% (comparable with GBM)

Sample ID	5' Gene	3' Gene	Discordant Read Pairs	Fusion Span Reads	Fusion Junction (s)	5' Gene Chr	3' Gene Chr	Validated ?
TCGA-AK-3456-01A-02R-1325-07	TFE3	SFPQ	175	129	1	chrX	chr1	Yes
TCGA-AK-3456-01A-02R-1325-07	SFPQ	TFE3	116	81	1	chr1	chrX	Yes
TCGA-A3-3313-01A-02R-1325-07	C6orf106	LRRC1	90	40	2	chr6	chr6	Yes
TCGA-A3-3313-01A-02R-1325-07	CYP39A1	LEMD2	37	9	1	chr6	chr6	Yes
TCGA-B2-4101-01A-02R-1277-07	FAM172A	FHIT	17	4	1	chr5	chr3	Yes
TCGA-AK-3445-01A-02R-1277-07	KIAA0802	LRRC41	14	6	1	chr18	chr1	Yes
TCGA-B0-5095-01A-01R-1420-07	GORASP2	WIPF1	14	2	1	chr2	chr2	Yes
TCGA-A3-3313-01A-02R-1325-07	ZNF193	MRPS18A	11	3	1	chr6	chr6	Yes
TCGA-A3-3313-01A-02R-1325-07	FTSJD2	GPX6	9	8	1	chr6	chr6	Yes
TCGA-B0-4945-01A-01R-1420-07	KIAA0427	GRM4	8	5	1	chr18	chr6	No
TCGA-B8-4143-01A-01R-1188-07	SLC36A1	TTC37	5	5	1	chr5	chr5	No

Gene fusion results in GBM

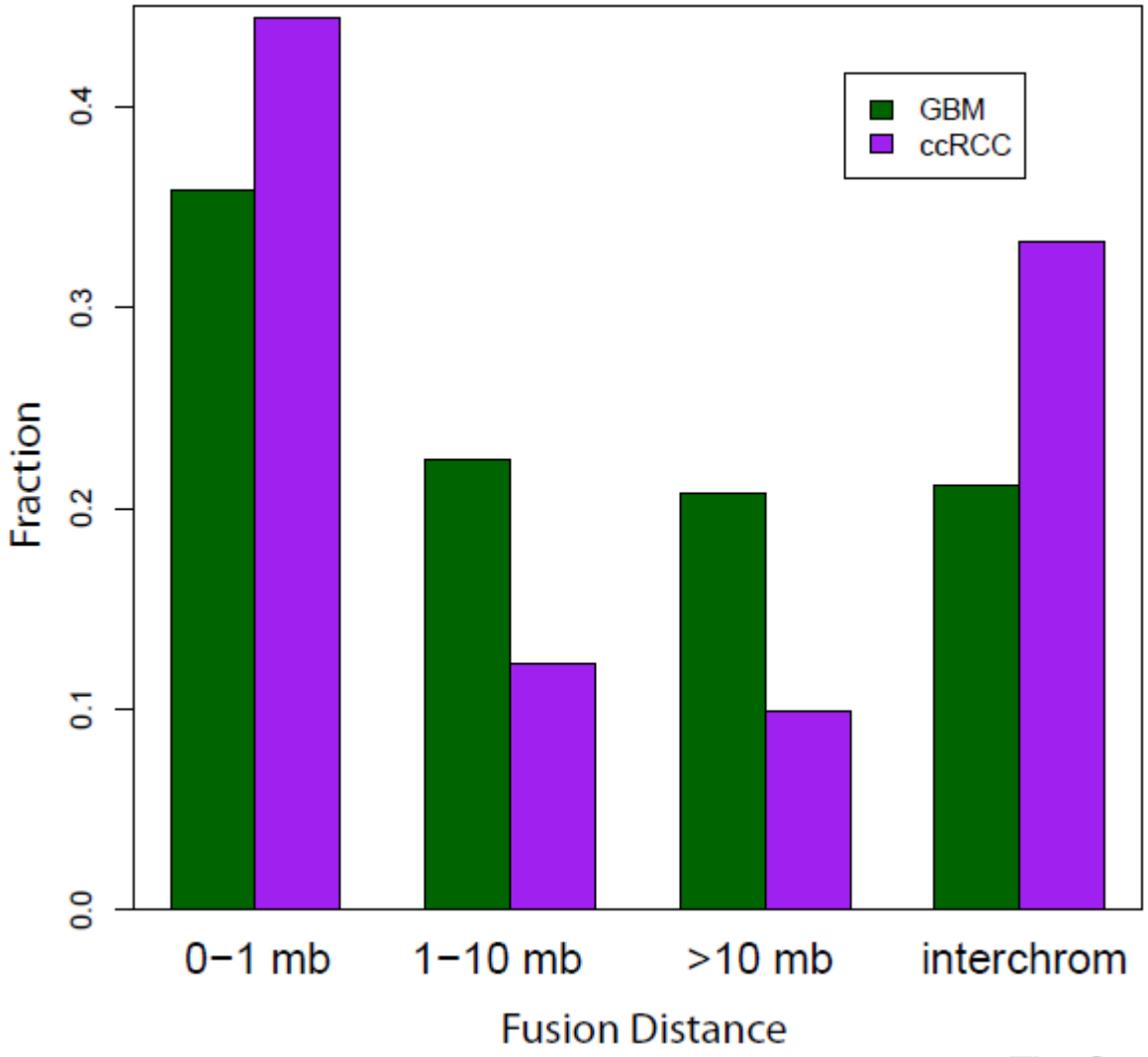
- 232 fusions were found in 164 GBMs, 68% of the samples have at least 1 fusion.
- Recurrent fusions:
  - *FGFR3-TACC3* (n=2, chr4) → Reported by Singh et al
  - *TFG-GPR128* (n=4, chr3) → Found in GBM & ccRCC
  - *EGFR fusions* (n=11, chr7) → Fusions in *EGFR* amplicon

Science

Transforming Fusions of *FGFR* and *TACC* Genes in Human Glioblastoma

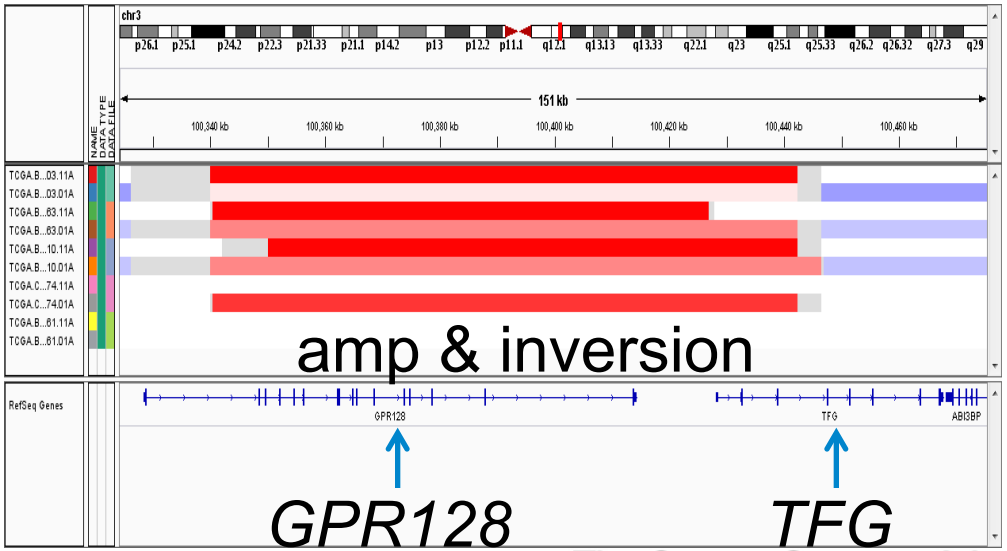
Devendra Singh,[1*] Joseph Minhow Chan,[2*] Pietro Zoppoli,[1*] Francesco Niola,[1*†] Ryan Sullivan,[1] Angelica Castano,[1] Eric Minwei Liu,[2] Jonathan Reichel,[2,3] Paola Porrati,[4] Serena Pellegatta,[4] Kunlong Qiu,[5] Zhibo Gao,[5] Michele Ceccarelli,[6] Riccardo Riccardi,[7] Daniel J. Brat,[8] Abhijit Guha,[9] Ken Aldape,[10] John G. Golfinos,[11] David Zagzag,[11,12] Tom Mikkelsen,[13] Gaetano Finocchiaro,[4] Anna Lasorella,[1,14,15]‡ Raul Rabadan,[2]‡ Antonio Iavarone[1,15,16]‡

Fusion distances in GBM and ccRCC



TFG-GPR128 fusion was evidenced by copy number from GBM and ccRCC

- Copy number variation in these two genes
  - previously annotated in:
    - a number of large human populations cohorts (see Database of Genomic Variant),
    - lymphoma and thyroid tissue tumors
    - As well as in healthy individuals
- Germline variant?

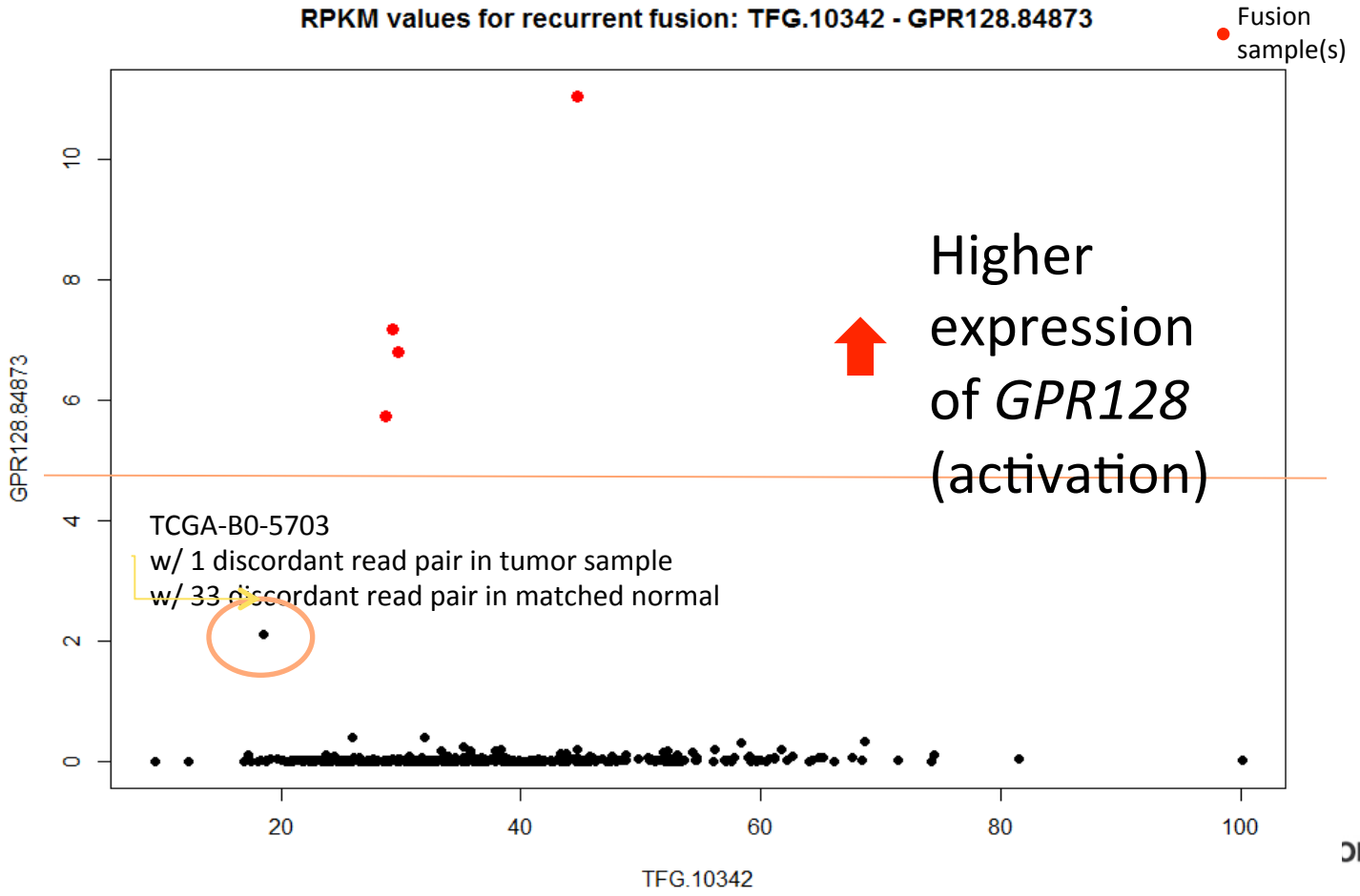


GUESS-ft searches for *TFG-GPR128* fusion in normal samples

- **GUESS-ft** was applied to all cancer samples with matched normal for *TFG-GPR128* fusion.
- The fusion was found in both tumor and matched normal in all cancer types examined:
  - Breast invasive carcinoma (n=106)
  - Kidney renal cell carcinoma (n=66)
  - Prostate adenocarcinoma (n=7)
- *TFG-GPR128* is a germline event.

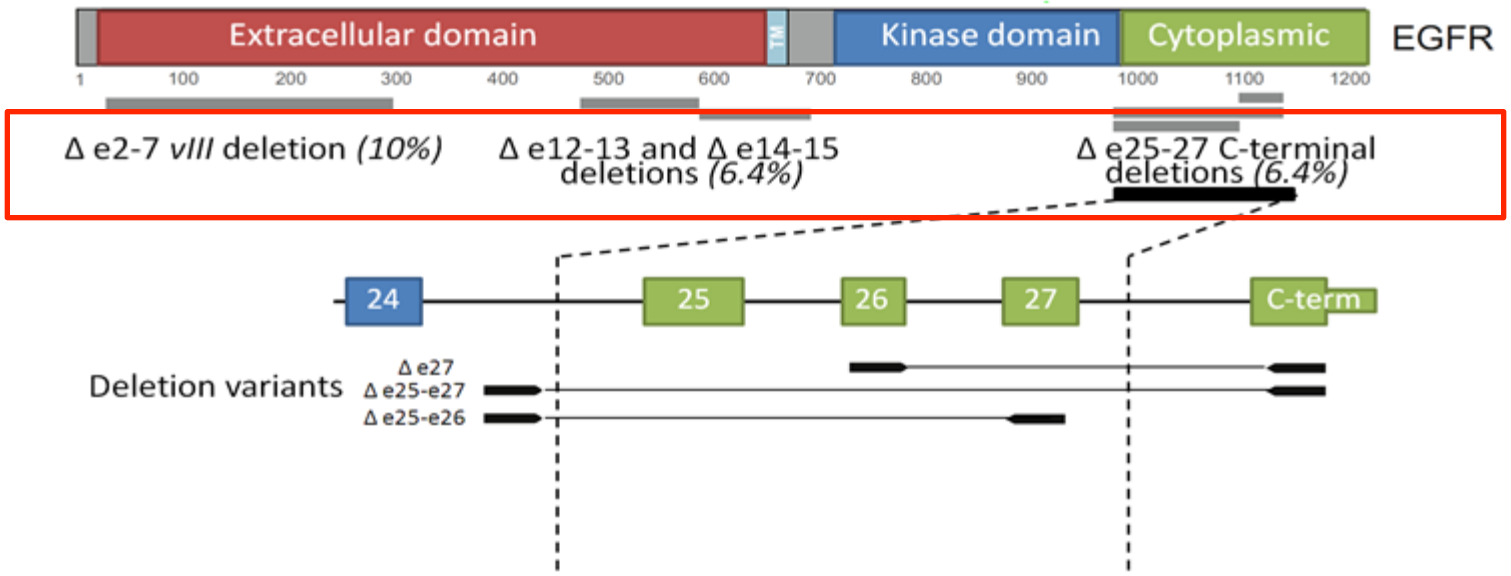
TFG-GPR128 fusion activates *GPR128* expression

- RPKM expression pattern seen in ccRCCs

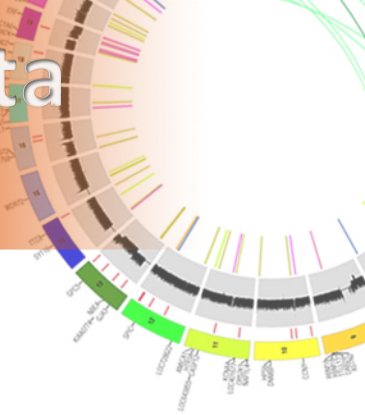


PRADA has a module "GUESS-ig" to identify intragenic rearrangements

- GUESS-ig: GUESS for ***i***ntra***g***enic rearrangements



Summary: *P*ipeline for *R*N*A*-seq *D*ata *A*nalysis



- PRADA HIGHLIGHTS:
  - Functionality (processing, QC&RPKM, fusion, etc).
  - Can be used as standalone version or within Portable Batch System (PBS) and Load Sharing Facility (LSF).
  - Modular steps allow users to flexibly "pause/resume" analysis, or run individual module.
  - Run samples in batch easily (~*180 samples in two weeks*).

PRADA is available in sourceforge (http://sourceforge.net/projects/prada/)

Home / Browse / Science & Engineering / Bio-Informatics / PRADA

Summary Files Reviews Wiki Blog Svn

PRADA

rahulsimham, roelverhaak, syzheng, wandaliztorres

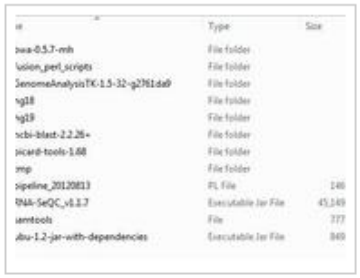
PRADA : Pipeline for RNA-Sequencing Data Analysis

Add a Review
3 Downloads (This Week)
Last Update: 2 hours ago

Download
PRADA.zip

Browse All Files

Tweet 0 +1 0 Like 0



Description

PRADA is designed for paired-end RNA sequencing data. PRADA focuses on the processing and analysis of mRNA profiles such as gene expression estimation and gene fusions identification. It implements several steps of analyses which are branched into two modules; (1) an initial module to process RNA-Seq data and calculate gene expression values, and (2) a module to identify gene fusion

ncer Genome Atlas

Acknowledgments

MD Anderson Cancer Center

- **Verhaak Lab**
- **Website:** www.virtuon.nl
  - **Roel Verhaak**
  - **Wandaliz Torres-Garcia**
  - **Rahul Vegesna**
  - Hoon Kim
  - Kosuke Yoshihara
  - Juan Emmanuel Martinez-Ledesma
  - Ji-Yeon Yang
- Rong Yao
- David Cogdell, Wei Zhang
- Zhiyong Ding, Peter German, Eric Jonasch

Collaborators

- Michael Berger (MSKCC)
- Andrey Shivachenko (BI)
- Gaddy Getz (BI)
- TCGA KIRC AWG
- TCGA GBM AWG

Poster #107