Batch correction & stuff

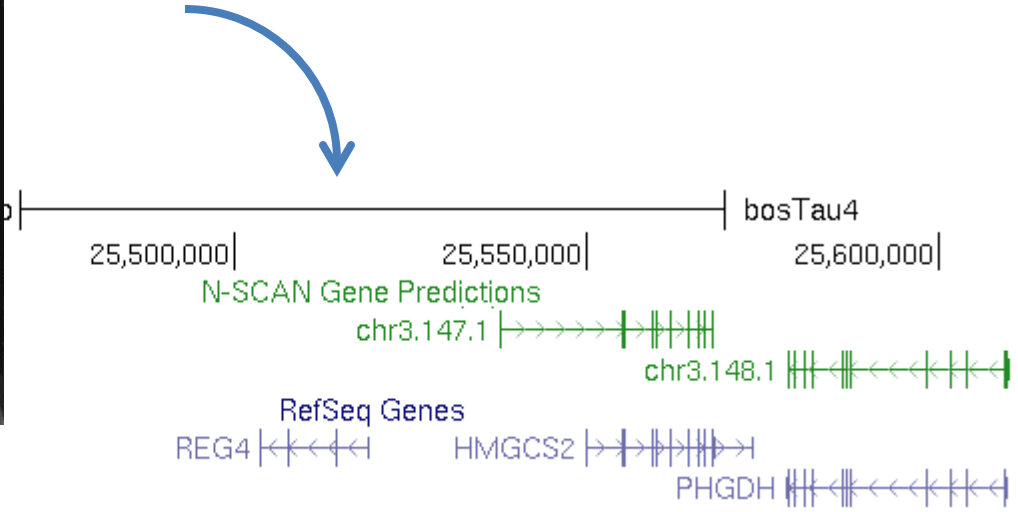
A tale of when not to use the Core Genomics Platform

Jeltje van Baren, Oct 18 2017

About me

- It’s pronounced ‘Yelcha’

Erasmus University, Rotterdam



WASHU

MBARI

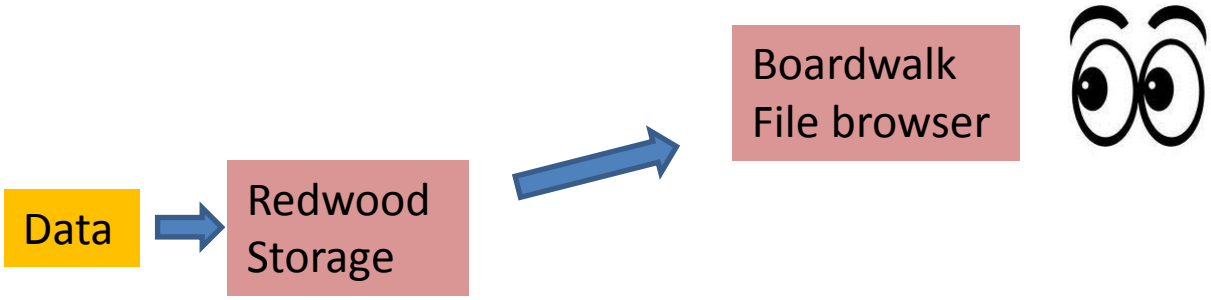
Today's talk

- Core Genomics Platform
- Batch effect correction
- WCDT fusion data

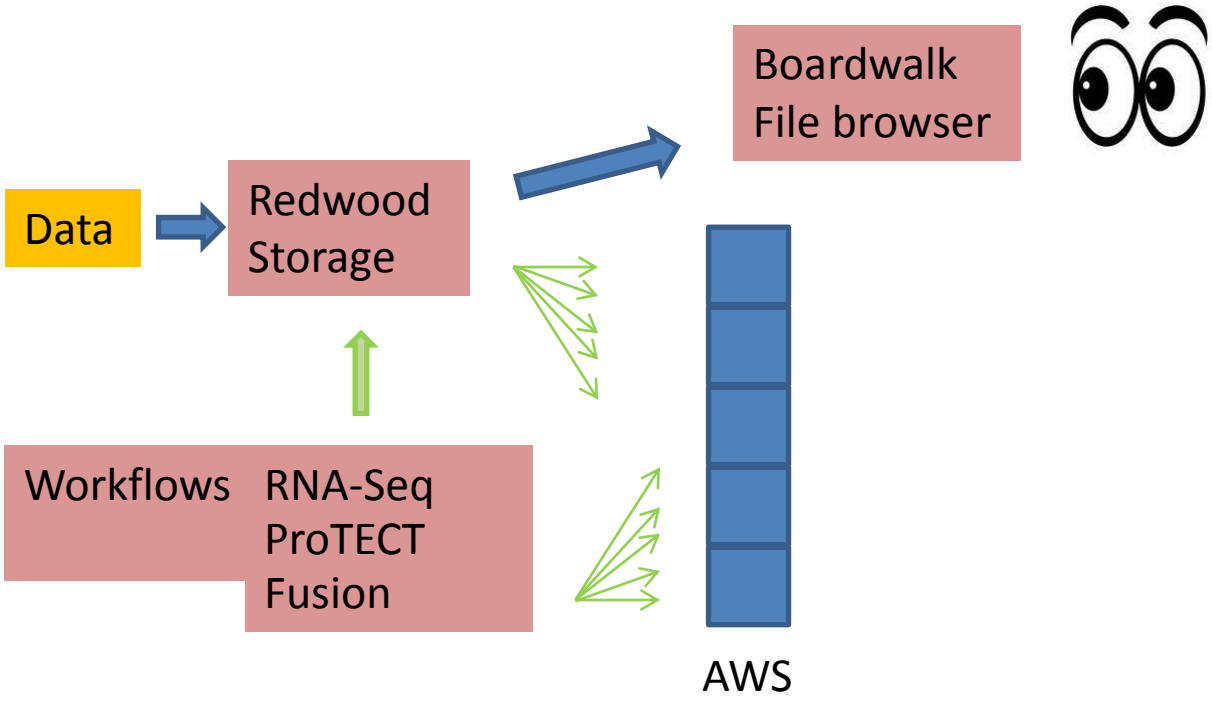
Core Genomics Platform

- Our mission as the Analysis Core is to help researchers in the UCSC Genomics Institute use modern tools to accurately and efficiently analyze large volumes of sequence data. We build the software that makes this possible (the Development team) and we also are responsible for running the "core" analysis pipelines for projects within the Institute (the Production team).

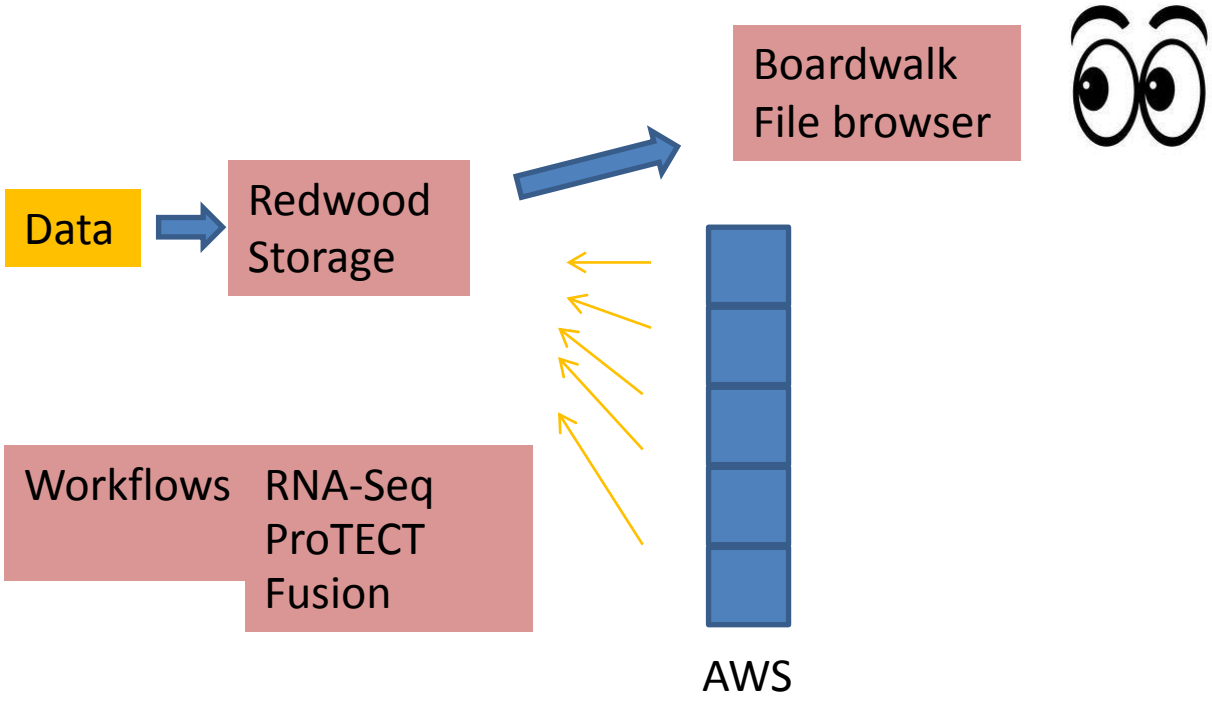
CGP



CGP


Boardwalk
File browser
Data
Redwood
Storage
Workflows
RNA-Seq
ProTECT
Fusion
AWS

CGP

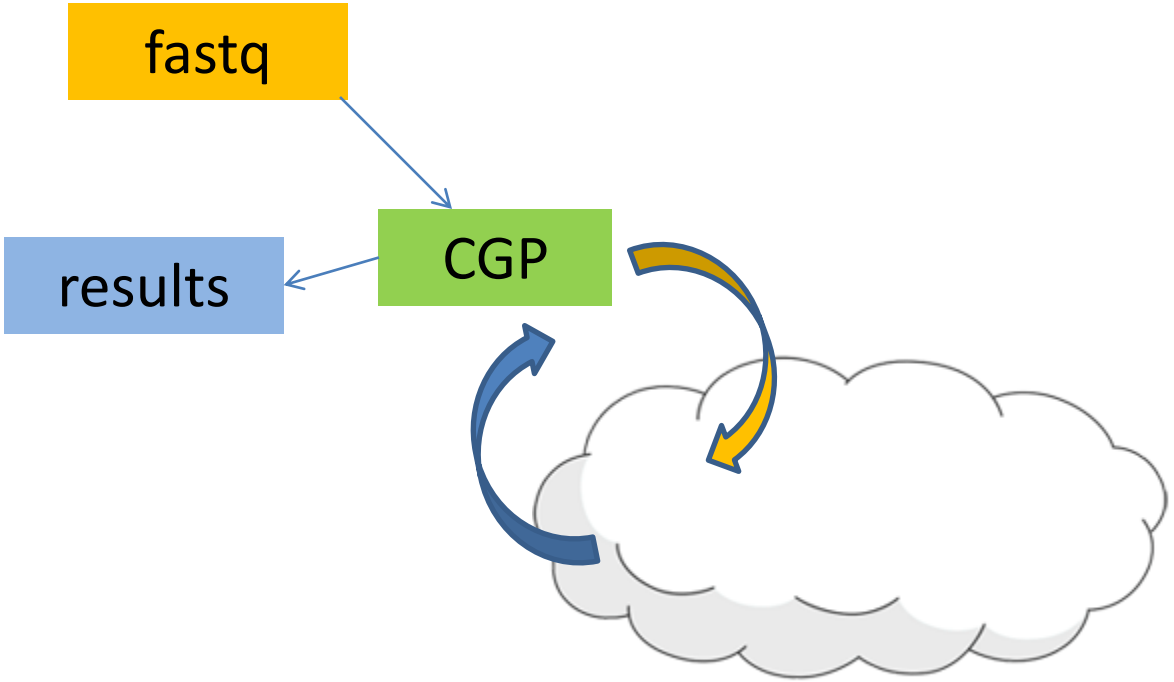

Boardwalk
File browser
Data
Redwood
Storage
Workflows
RNA-Seq
ProTECT
Fusion
AWS

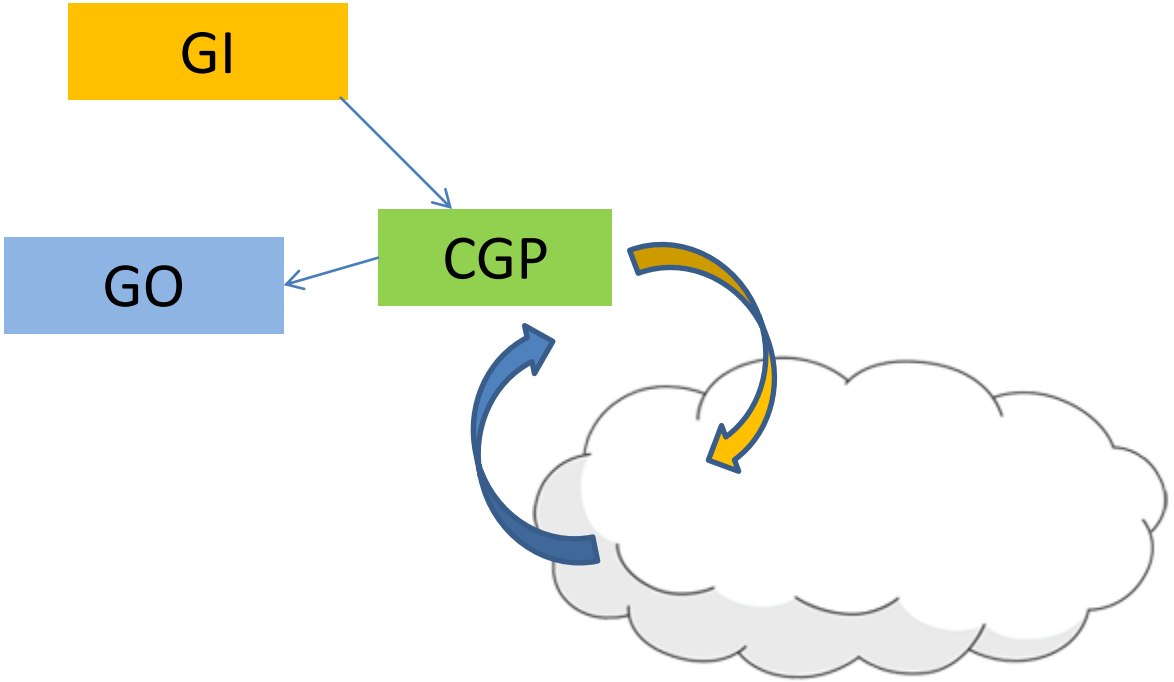
CGP Workflows are mostly Docker based

- Docker images are frozen VMs
- Stable environment
- Easy to set up
- If you can run it on your Linux machine, you can make a docker image

CGP is Great for

- Large datasets
- Precisely defined workflows
  - Sequence alignment
  - Quantification
  - Fusion
- Reproducability

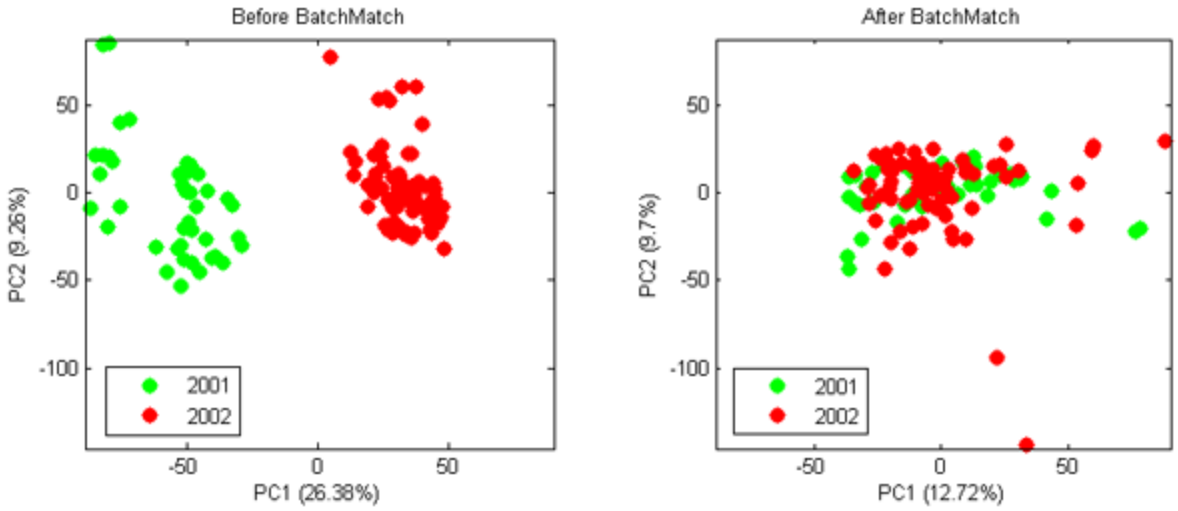

fastq
CGP
results


GI
CGP
GO

CGP is not so great for

- Interactive workflows (QC!)
- Distributed workflows (BCBio)
- Replacing science
  - Don’t let it make decisions for you
- Batch analysis
  - Can’t add additional data later

Batch effect correction



"Let's run some program to fix this"

(Yulia's) combat.R

- File: gene by sample matrix
- File: sample ID and batch

Magic happens

Reason 1 to not put in Docker: Metadata is *messy*.

Combat doesn’t like zeroes

	Batch1	Batch1	Batch1	Batch2	Batch2	Batch2
Gene1	0	0	0	0	0	0
Gene2	0	0	0	0	23	19
Gene3	0	34	0	0	0	0
Gene4	15	0	2	16	12	21

Jaffe et al. BMC Bioinformatics (2015) 16:372
DOI 10.1186/s12859-015-0808-5

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Practical impacts of genomic data "cleaning" on biological discovery using surrogate variable analysis

Andrew E. Jaffe[1,2], Thomas Hyde[1,3,4], Joel Kleinman[1,3], Daniel R. Weinbergern[1,3,4,5,6], Joshua G. Chenoweth[1], Ronald D. McKay[1], Jeffrey T. Leek[2] and Carlo Colantuoni[1,3,5*]

Abstract

Background: Genomic data production is at its highest level and continues to increase, making available novel primary data and existing public data to researchers for exploration. Here we explore the consequences of "batch" correction for biological discovery in two publicly available expression datasets. We consider this to include the estimation of and adjustment for wide-spread systematic heterogeneity in genomic measurements that is unrelated to the effects under study, whether it be technical or biological in nature.

Methods: We present three illustrative data analyses using surrogate variable analysis (SVA) and describe how to perform artifact discovery in light of natural heterogeneity within biological groups, secondary biological questions of interest, and non-linear treatment effects in a dataset profiling differentiating pluripotent cells (GSE32923) and another from human brain tissue (GSE30272).

PMID 26545828

Jaffe *et al. BMC Bioinformatics* (2015) 16:372
DOI 10.1186/s12859-015-0808-5

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Practical impacts of genomic data "cleaning" on biological discovery using

Batch correction alters expression values – DO NOT USE for differential expression analysis

primary data and existing public data to researchers for exploration. Here we explore the consequences of "batch" correction for biological discovery in two publicly available expression datasets. We consider this to include the estimation of and adjustment for wide-spread systematic heterogeneity in genomic measurements that is unrelated to the effects under study, whether it be technical or biological in nature.

Methods: We present three illustrative data analyses using surrogate variable analysis (SVA) and describe how to perform artifact discovery in light of natural heterogeneity within biological groups, secondary biological questions of interest, and non-linear treatment effects in a dataset profiling differentiating pluripotent cells (GSE32923) and another from human brain tissue (GSE30272).

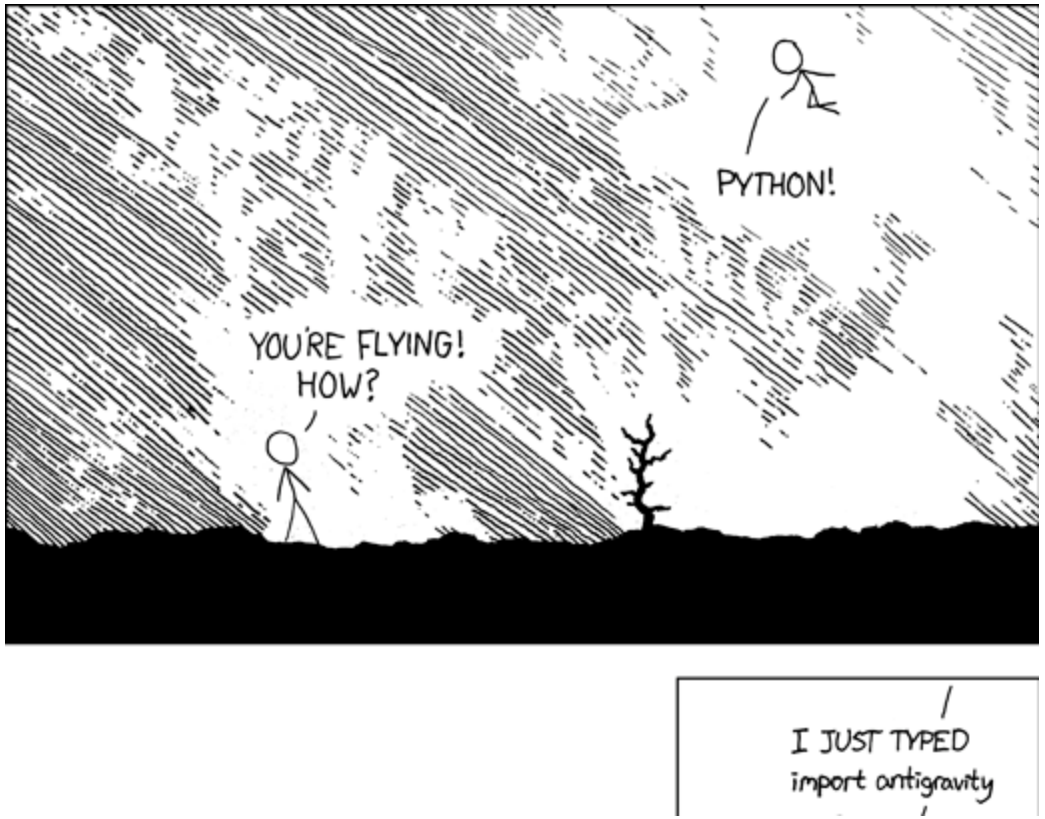
PMID 26545828

Avoiding GIGO

- Plot your data!

Avoiding GIGO

- Plot your data!



…using R

BatchQC is interactive

```
source("https://bioconductor.org/biocLite.R")
biocLite(c("BatchQC"))
library(BatchQC)

metadf <- read.tsv("./testdata/rnaseq_annotations.tsv“)
exprdata <-
read.tsv("./testdata/data_normalized_sample_names.tab")
batch = metadf$Sequencer
condition = metadf$HistoGroup

batchQC(dat=exprdata, batch=batch, condition=condition,
report_file="batchqc_report.html", report_dir=".",
report_option_binary="111111111",
view_report=TRUE, interactive=TRUE, batchqc_output=TRUE)
```

BatchQC is interactive

```
source("https://bioconductor.org/biocLite.R")
biocLite(c("BatchQC"))
library(BatchQC)

metadf <- read.tsv("./testdata/rnaseq_annotations.tsv“)
exprdata <-
read.tsv("./testdata/data_normalized_sample_names.tab")
batch = metadf$Sequencer
condition = metadf$HistoGroup

batchQC(dat=exprdata, batch=batch, condition=condition,
report_file="batchqc_report.html", report_dir=".",
report_option_binary="111111111",
view_report=TRUE, interactive=TRUE, batchqc_output=TRUE)
```

- Must filter out 0-values from exprdata
- Must remove samples with no $HistoGroup
- Must downsample for visualization

- http://report.html

Fusion finding

- Treehouse implementation of STAR-Fusion
- Available on CGP
- Ran on all available RNA-Seq WCDT samples

TMPRSS2–ERG fusion co-opts master transcription factors and activates NOTCH signaling in primary prostate cancer

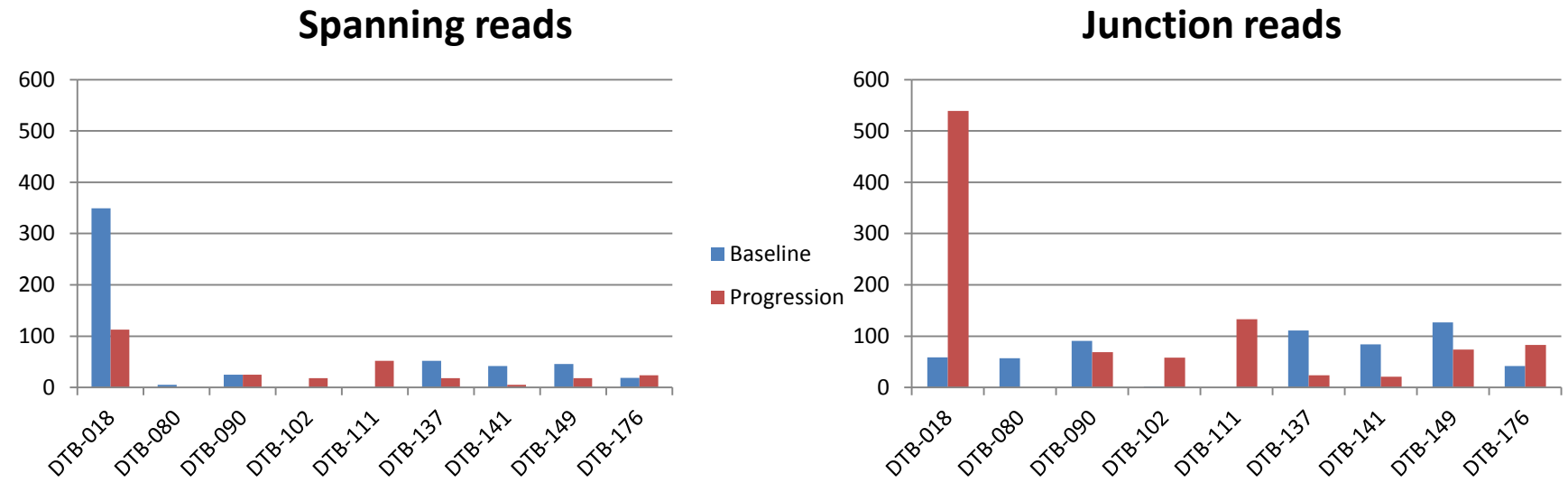
Ken J Kron[1,7], Alexander Murison[1,7], Stanley Zhou[1,2], Vincent Huang[3], Takafumi N[illegible]naguchi[3], Yu-Jia Shiah[3], Michael Fraser[1], Theodorus van der Kwast[4], Paul C Boutros[2,3,5], R[illegible]t G Bristow[1,2,6] & Mathieu Lupien[1–3]

TMPRSS2–ERG (T2E) structural rearrangements typify ~50% of prostate tumors an[illegible]erexpression of the ERG transcription factor. Using chromatin, genomic and expression data, we show distinct [illegible]latory landscapes between T2E-positive and non-T2E primary prostate tumors, which include clusters of regulatory [illegible]ents (COREs). This difference is mediated by ERG co-option of HOXB13 and FOXA1, implementing a T2E-specific transcriptional profile. We also report a T2E-specific CORE on the structurally rearranged *ERG* locus arising from spreading of the *TMPRSS2* locus pre-existing CORE, assisting in its overexpression. Finally, we show that the T2E-specific *cis*-regulatory landscape underlies a vulnerability against the NOTCH pathway. Indeed, NOTCH pathway inhibition antagonizes the growth and invasion of T2E-positive prostate cancer cells. Taken together, our work shows that overexpressed ERG co-opts master transcription factors to deploy a unique *cis*-regulatory landscape, inducing a druggable dependency on NOTCH signaling in T2E-positive prostate tumors.

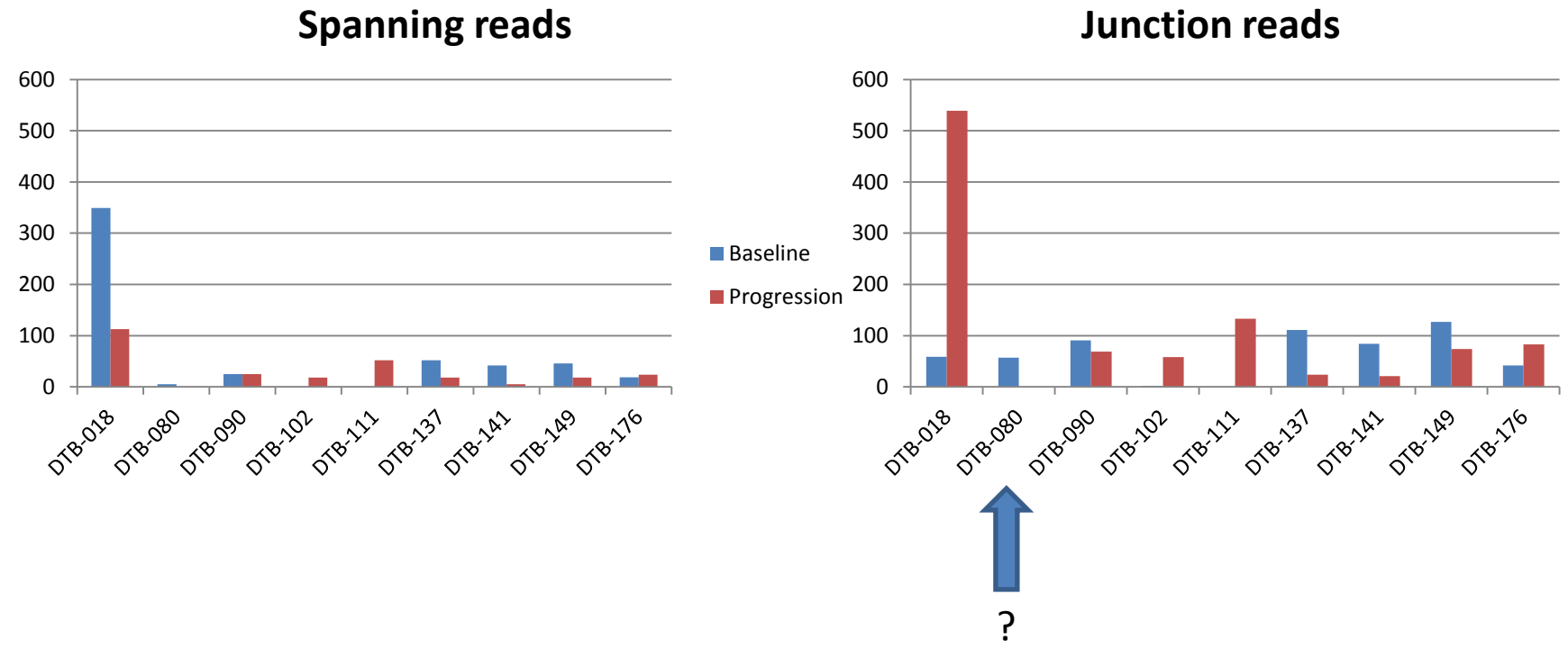
Nat Genet. 2017 Sep;49(9):1336-1345

TMPRSS2-ERG occurs as a breakpoint in 29 of 95 Baseline samples

Two kinds of fusion reads



Two kinds of fusion reads



“Novel” breakpoint C15orf57-CBX3

	Junction	Spanning		Junction	Spanning
DTB-011_Baseline	3	1	DTB-090_Progression	3	0
DTB-069_Baseline	4	1	DTB-102_Progression	4	0
DTB-104_Baseline	3	0	DTB-194_Progression	9	9
DTB-120_Baseline	3	1			
DTB-127_Baseline	4	2			
DTB-174_Baseline	13	6			
DTB-190_Baseline	3	1			

- CBX3 promotes colon cancer cell proliferation by CDK6 kinase
- C15orf57 highly expressed in EBV transformed lymphocytes