Exome-seq Analysis
The CCBR's perspective

Li Jia
CCR Collaborative Bioinformatics Resource (CCBR)
03/18/2015

Outlines

- Lesson learned from the bad experimental design
- Best practice in CCBR
- Functional annotation
- How to collaborate with CCBR – guide to success

Outlines

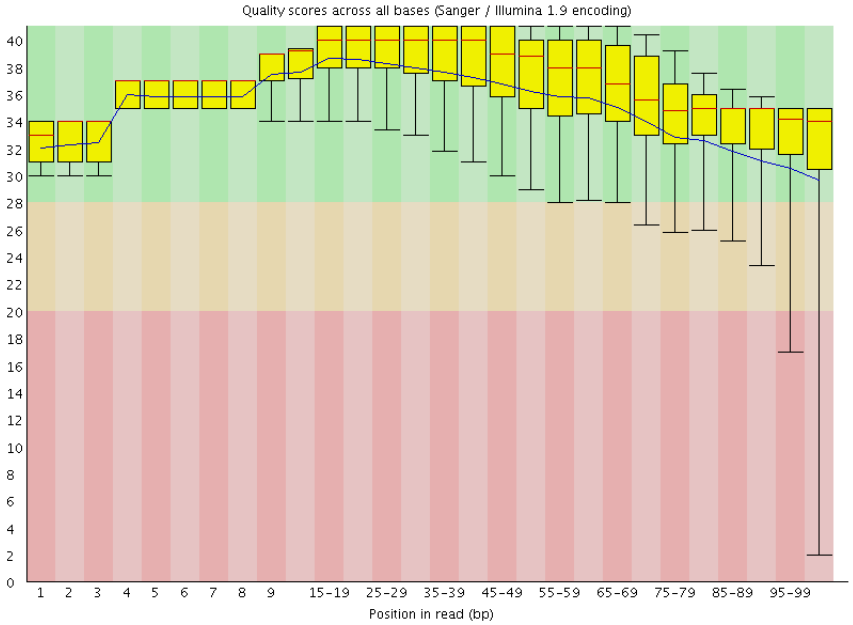
- Lesson learned from the bad experimental design
- Best practice in CCBR
- Functional annotation
- How to collaborate with CCBR – guide to success

Experimental design

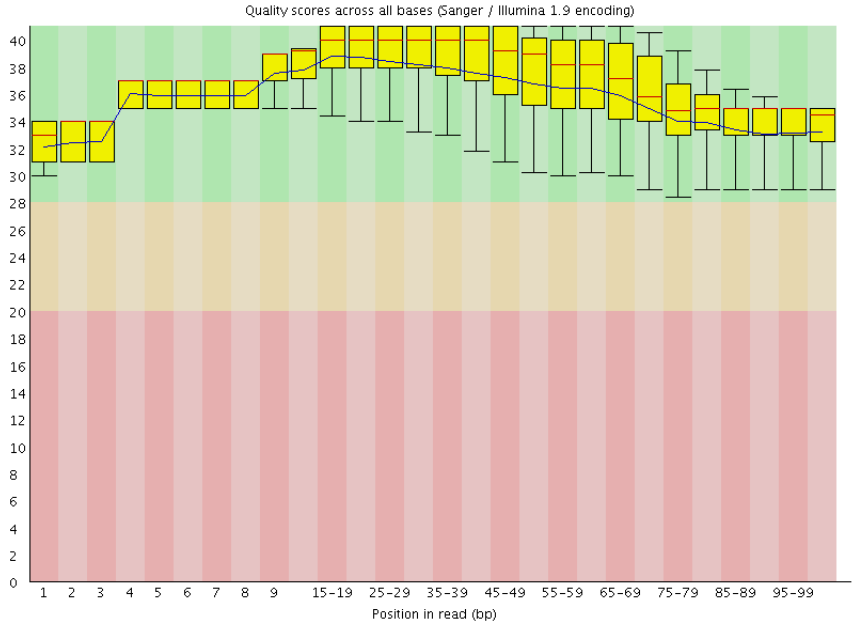
- Goal: Identification of mutations that are shared by all KO tumor samples
- Samples:
  - Cases: gene-deficient B cell lymphomas (10 samples)
  - Controls: one is from 129 strains and another is gene KO T cell-deficient mouse (2 samples)
- Exome-Seq
  - Compare cases and controls

QC

- FastQC & trimming of reads

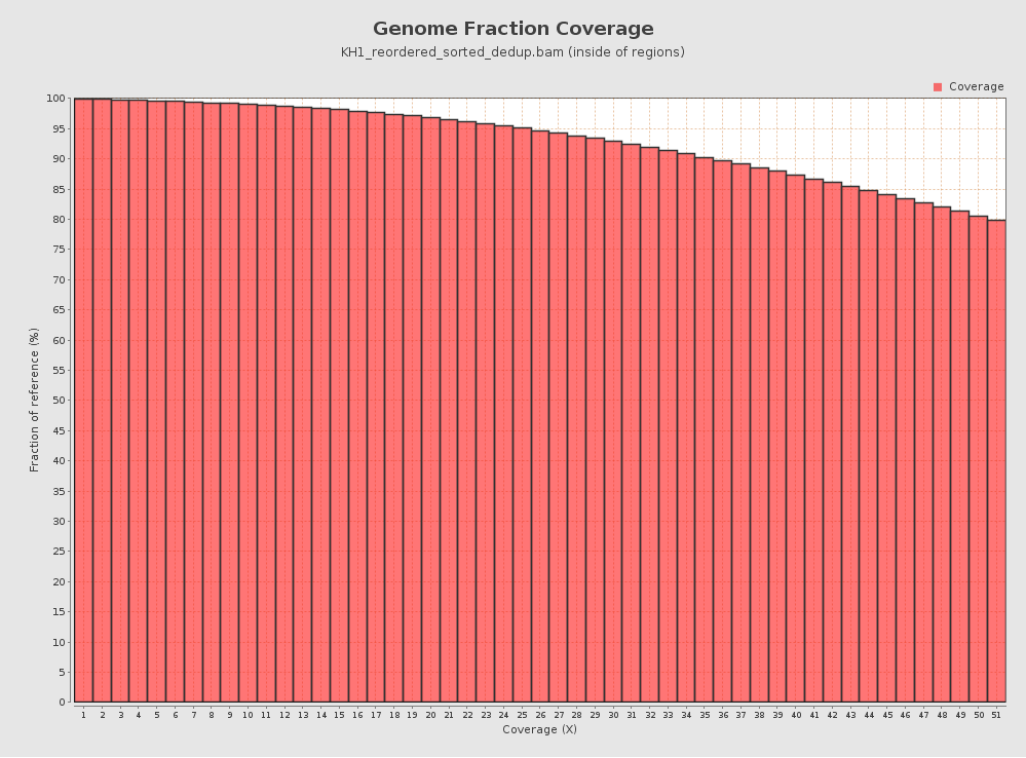


Raw fastq



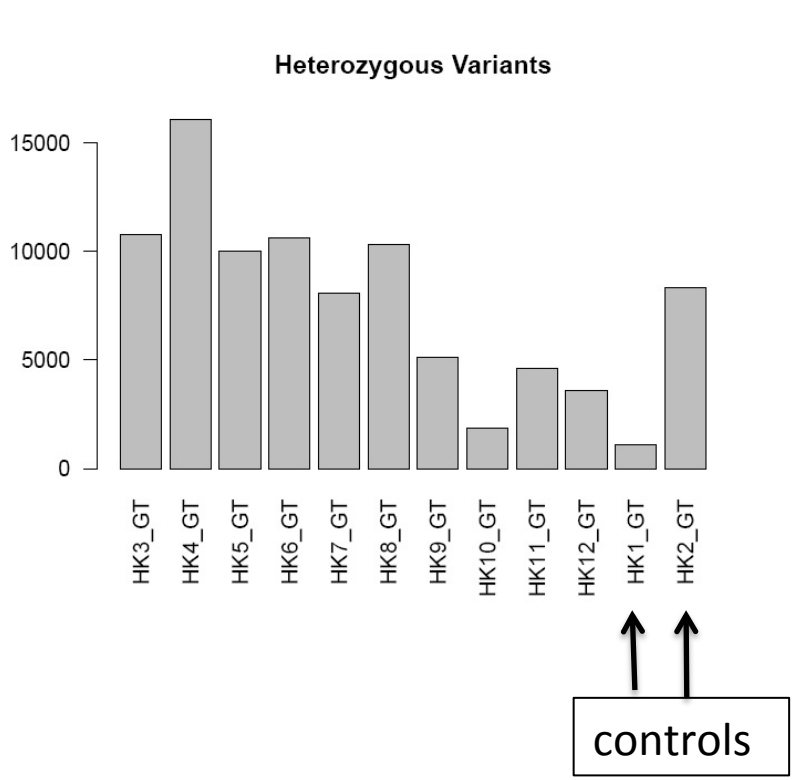
Trimmed fastq

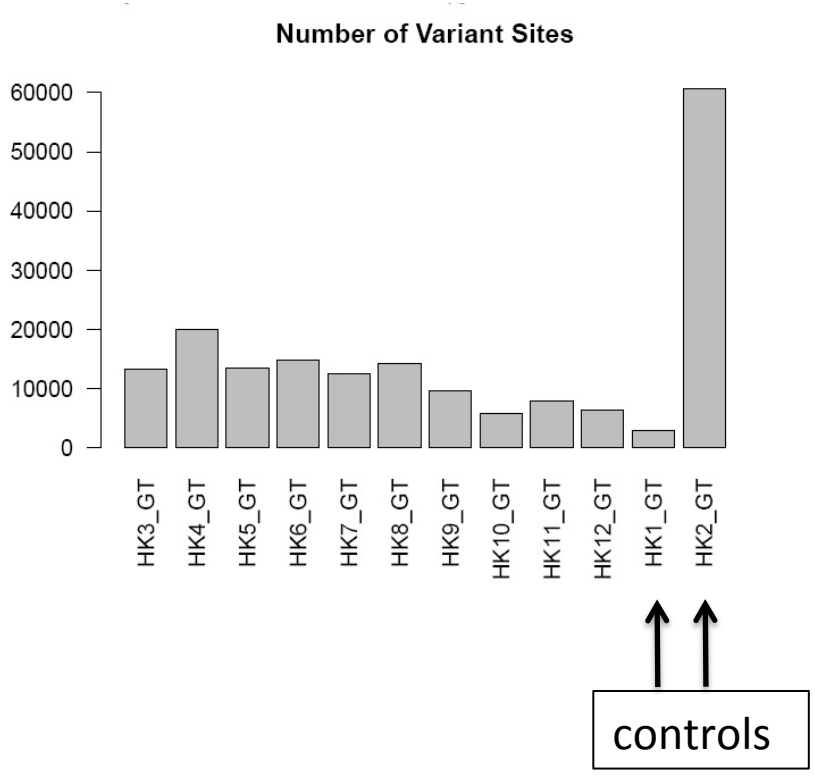
Coverage analysis



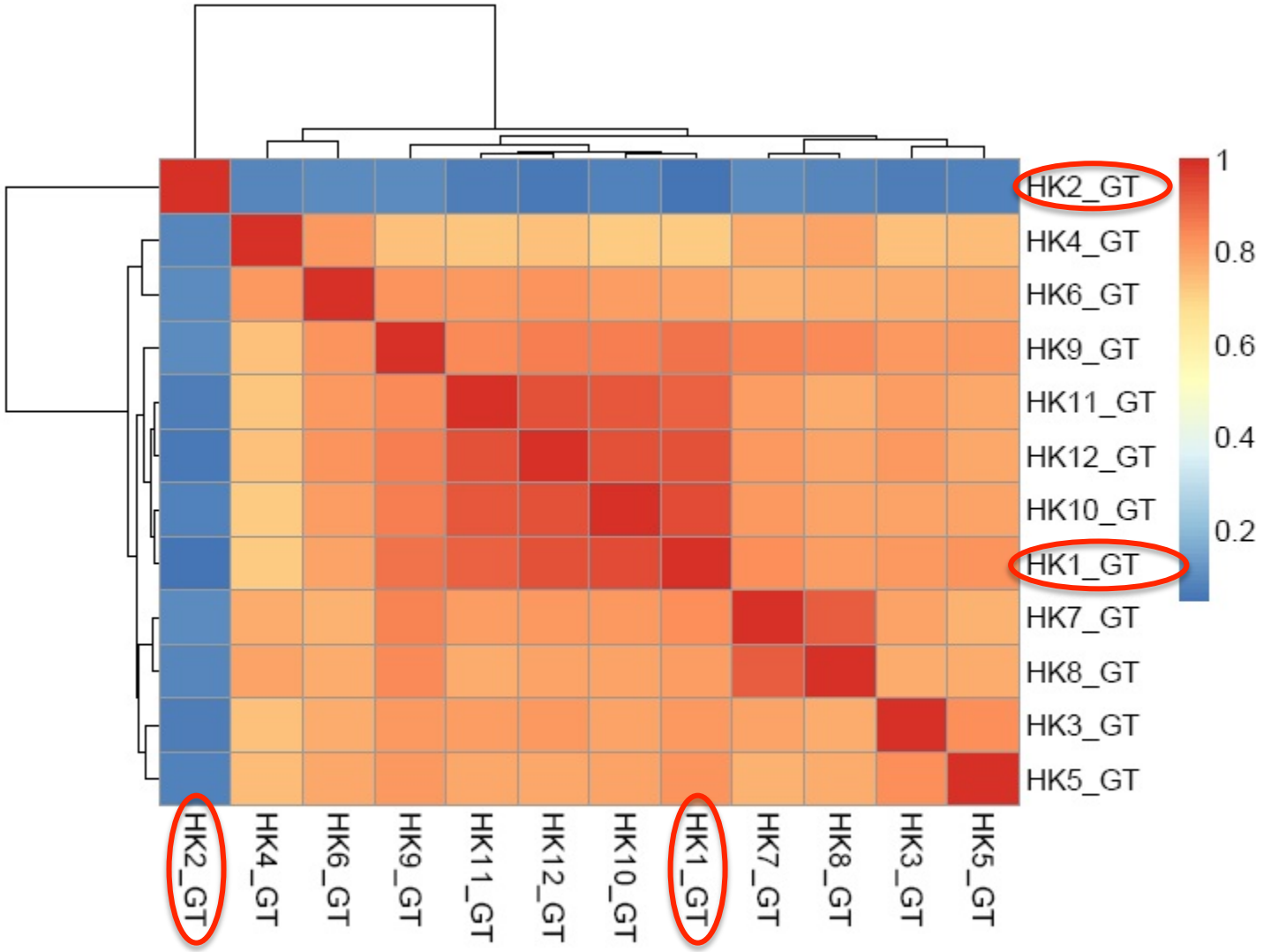
	Mean	>= 10x(%)	>=20x(%)
KH10	144.57	99.05	96.6%
KH12	135.67	97.74	93.49
KH11	146.72	99.06	96.71
KH1	147.79	99.17	97.12
KH3	172.94	99.2	97.45
KH2	128	98.2	95
KH4	133.21	98.56	95.22
KH8	138.22	98.86	96
KH5	145.21	98.98	96.45
KH9	131.69	98.95	96.22
KH6	153.8	98.81	96.47
KH7	147.18	98.96	96.74

Barplot of the Counts



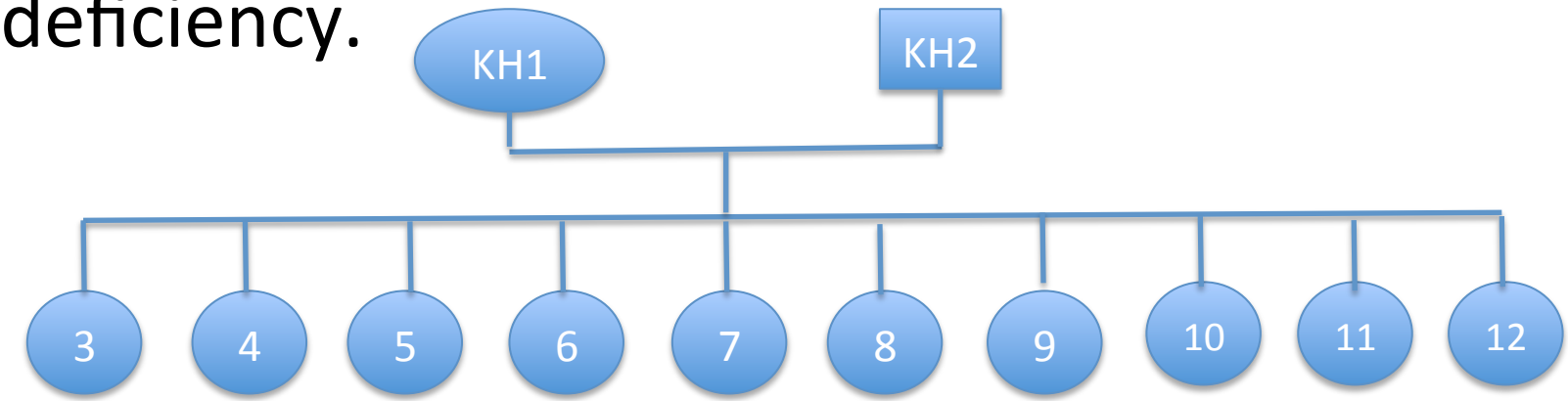


Heatmap of the Samples



Recommendations

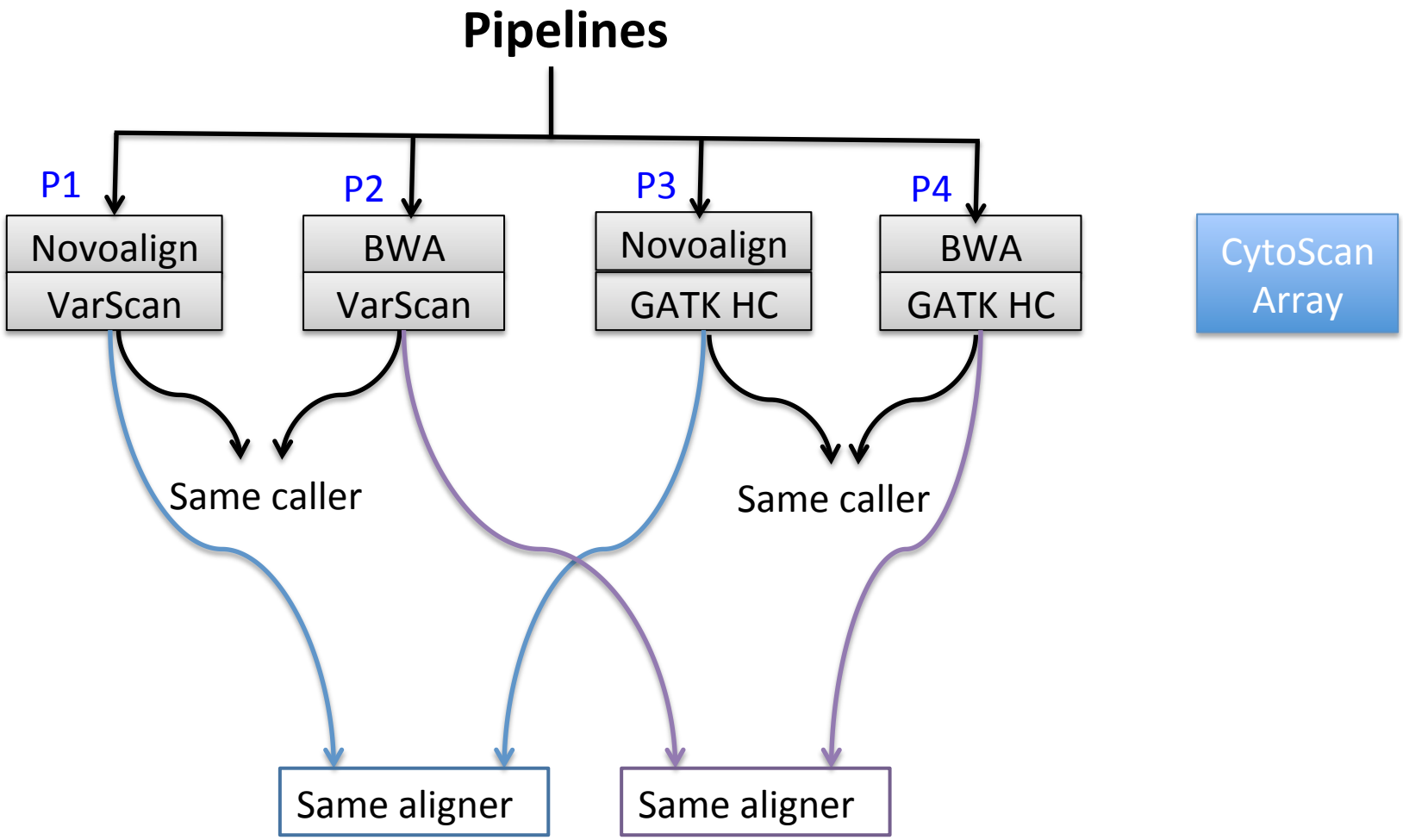
- The control samples in the experiments should use the same background
- Case-control study needs to have case-control pair, not 2 controls and 10 cases. It is also not germline mutation, the cases use some genes deficiency.



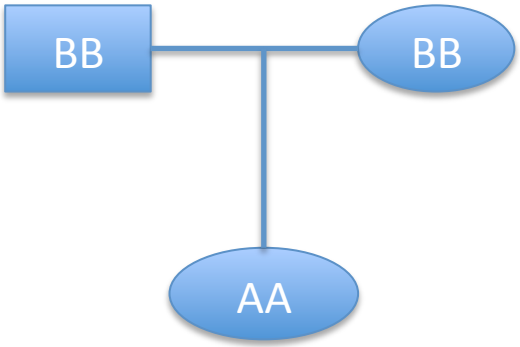
Outlines

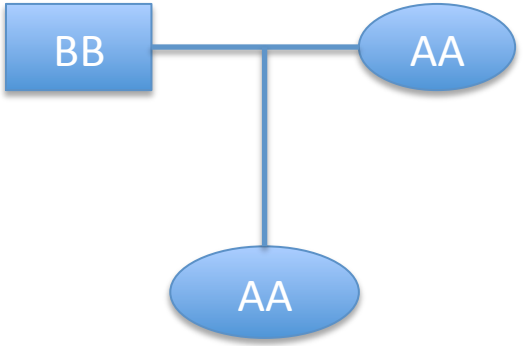
- Lesson learned from the bad experimental design
- Best practice in CCBR
- Functional annotation
- How to collaborate with CCBR – guide to success

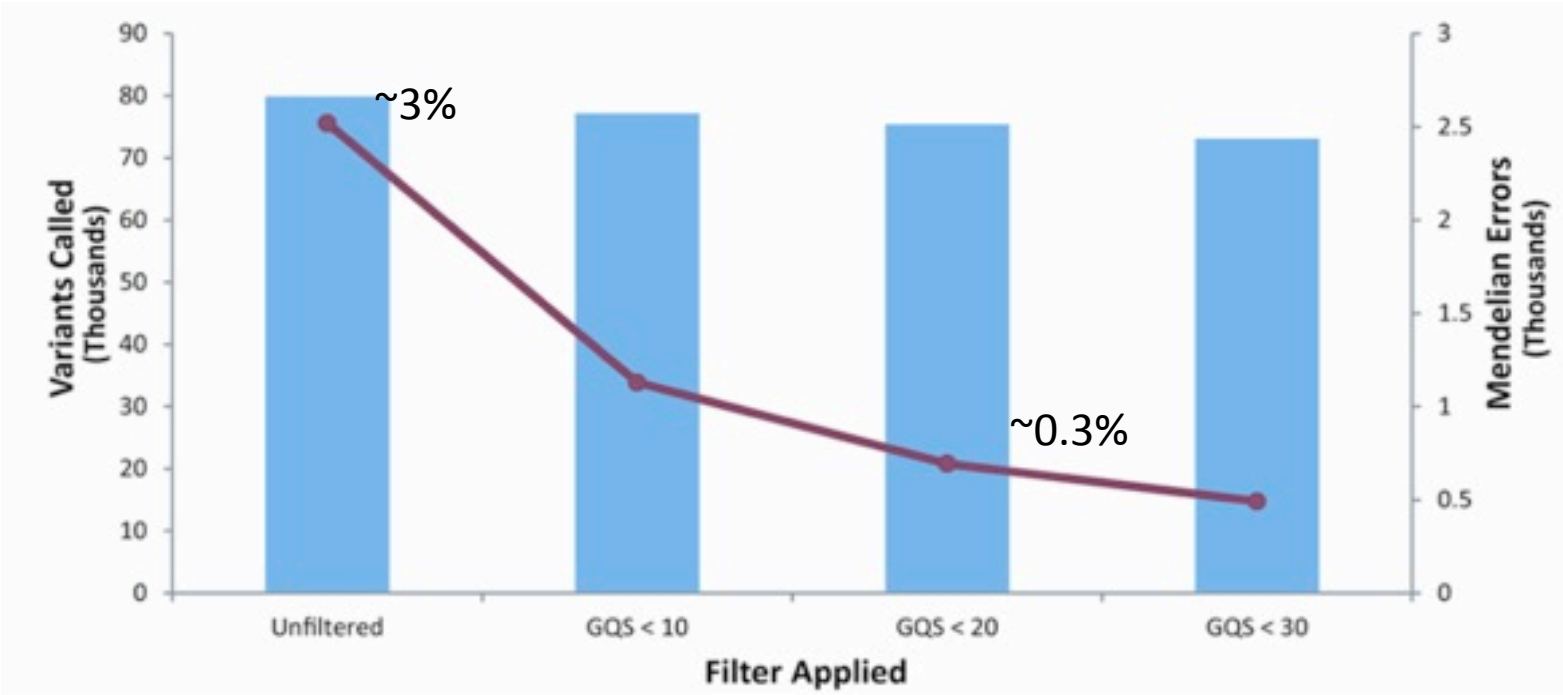
Pipelines in Germline



Mendelian Error







Comparisons of the Pipelines

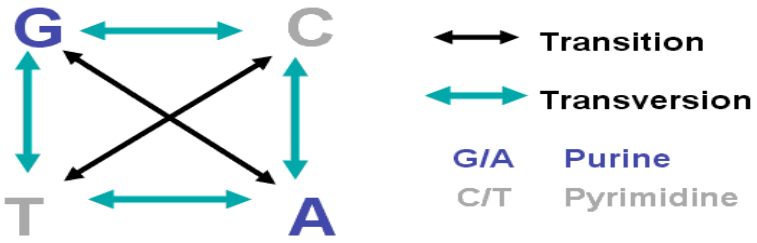
	P1:Novo-VS	P2: BWA-VS	P3:Novo-HC	P4:BWA-HC
% Mendel errors	5.2	5.81	0.81	0.99
Ti/Tv	2.28	2.43	2.45	2.44
% overlap with Cytoscan call	93.01%	92.71%	93.57%	93.47%

Note:

Ti/Tv on raw variants

Expected Ti/Tv for human exome : 2.4-2.8

BWA alignment used trimmed fastq files



Comparison of BWA and Novoalign

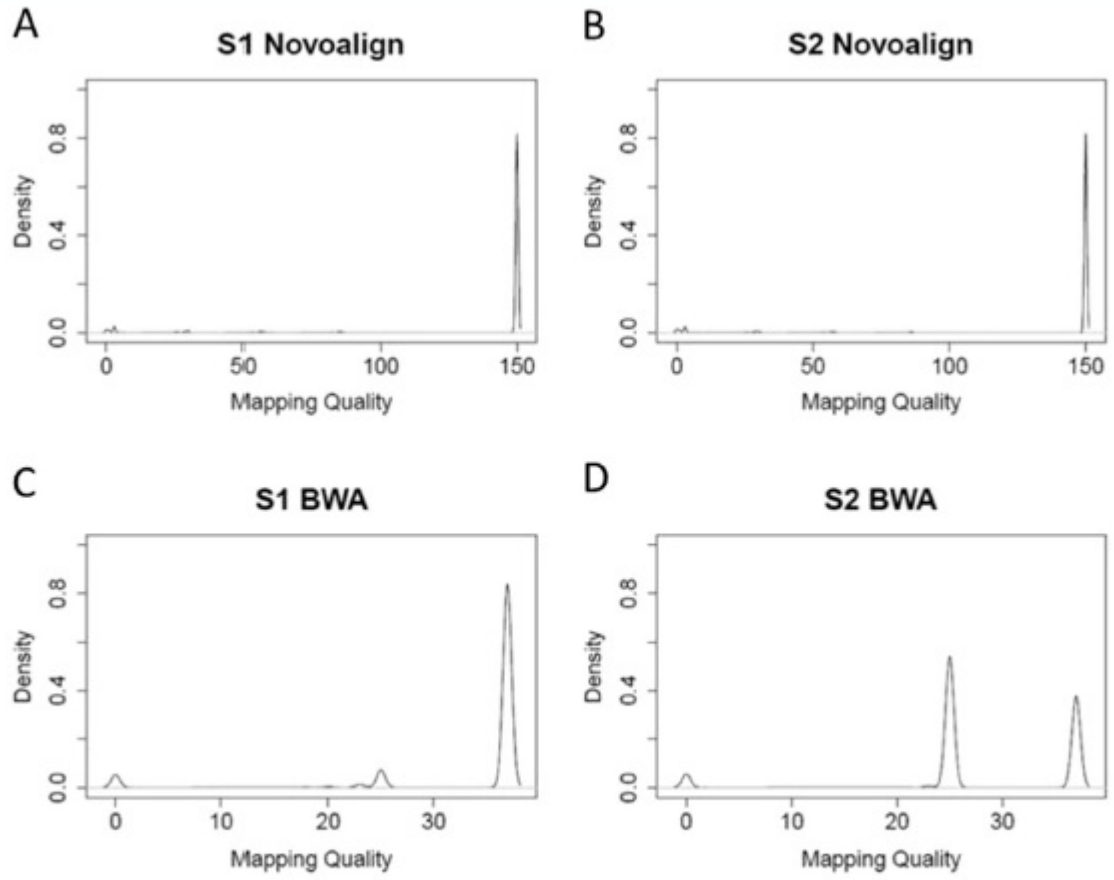
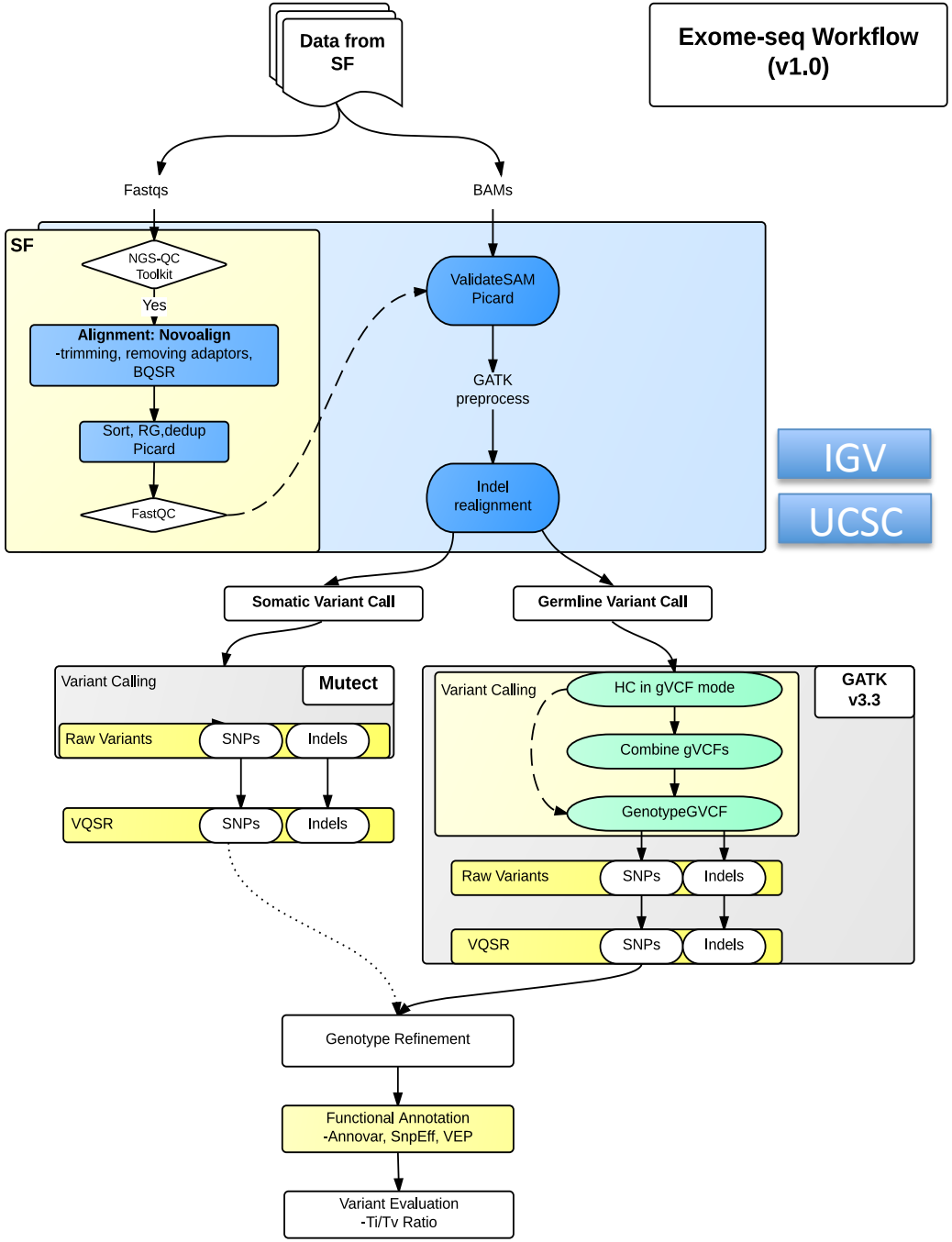
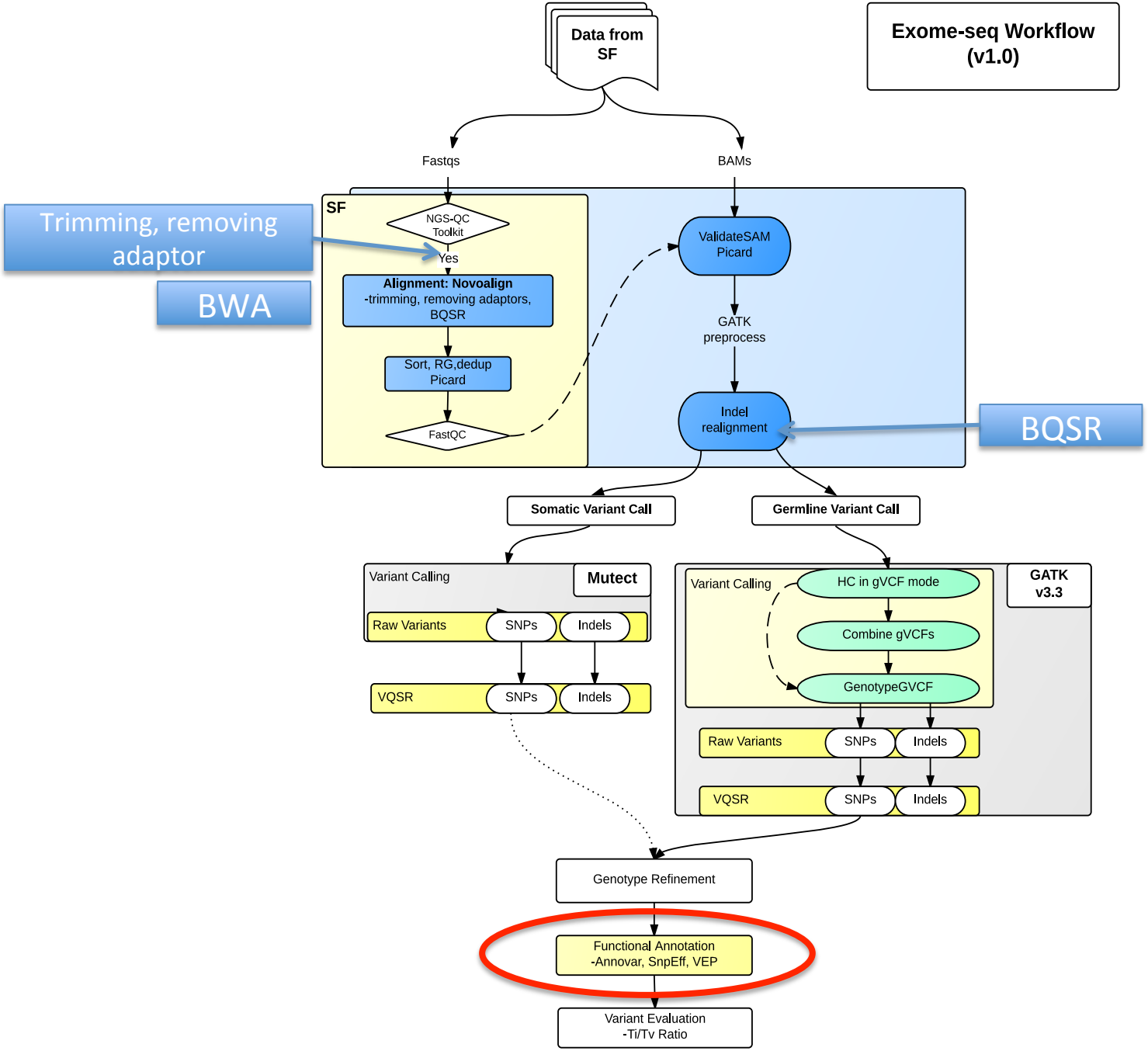


Figure 3 Mapping quality scores reported in Novoalign and BWA. Alignment is performed on both the untrimmed S1 and S2 data sets, with one alignment randomly reported for each read.

Yu et al. BioData Mining 2012, 5:6


Exome-seq Workflow
(v1.0)
Data from
SF
Fastqs
BAMs
SF
NGS-QC
Toolkit
Yes
Alignment: Novoalign
-trimming, removing adaptors,
BQSR
Sort, RG,dedup
Picard
FastQC
ValidateSAM
Picard
GATK
preprocess
Indel
realignment
IGV
UCSC
Somatic Variant Call
Germline Variant Call
Variant Calling
Mutect
Raw Variants
SNPs
Indels
VQSR
SNPs
Indels
Variant Calling
HC in gVCF mode
Combine gVCFs
GenotypeGVCF
GATK
v3.3
Raw Variants
SNPs
Indels
VQSR
SNPs
Indels
Genotype Refinement
Functional Annotation
-Annovar, SnpEff, VEP
Variant Evaluation
-Ti/Tv Ratio


Exome-seq Workflow
(v1.0)
Data from SF
Fastqs
BAMs
Trimming, removing adaptor
BWA
SF
NGS-QC Toolkit
Yes
Alignment: Novoalign
-trimming, removing adaptors, BQSR
Sort, RG,dedup Picard
FastQC
ValidateSAM Picard
GATK preprocess
Indel realignment
BQSR
Somatic Variant Call
Germline Variant Call
Variant Calling
Mutect
Raw Variants
SNPs
Indels
VQSR
SNPs
Indels
Variant Calling
HC in gVCF mode
Combine gVCFs
GenotypeGVCF
GATK v3.3
Raw Variants
SNPs
Indels
VQSR
SNPs
Indels
Genotype Refinement
Functional Annotation
-Annovar, SnpEff, VEP
Variant Evaluation
-Ti/Tv Ratio

Functional Annotation

- ANNOVAR
- SnpEff
- VEP – Variant Effect Predictor

What is the effect?

In a gene?

Amino-acid change?

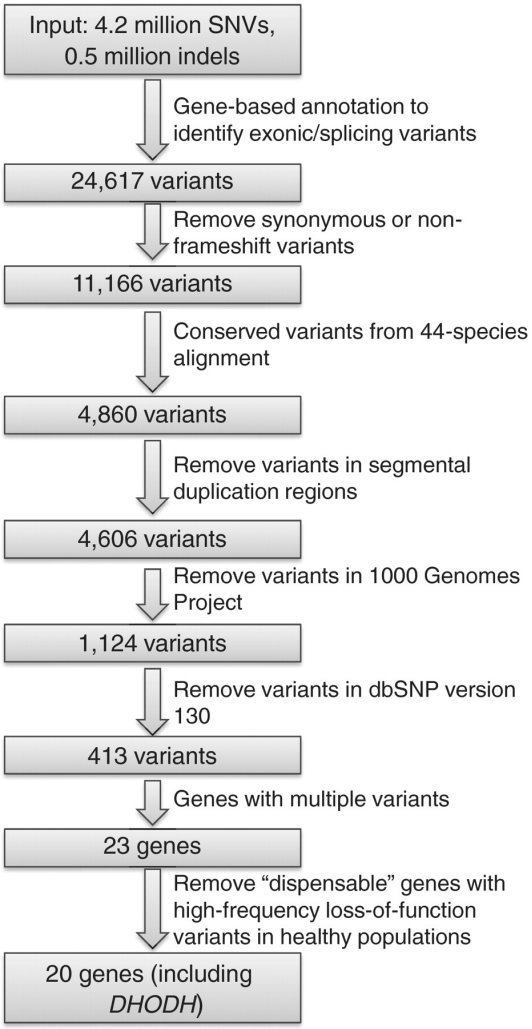
Coding?

Detrimental for function?

Annotation Tools

- Annovar
  - Exonic splicing, distance to nearest gene, indels
  - Local scripts
  - Create your own databases
- SnpEff
  - Integration with GATK and Galaxy, read and write VCF format
  - Local Java program
- VEP – Variant Effect Predictor
  - Include more comprehensive mouse genome databases

Identification of genes responsible for Miller syndrome using a synthetic data set.



Kai Wang et al. Nucl. Acids Res. 2010;38:e164

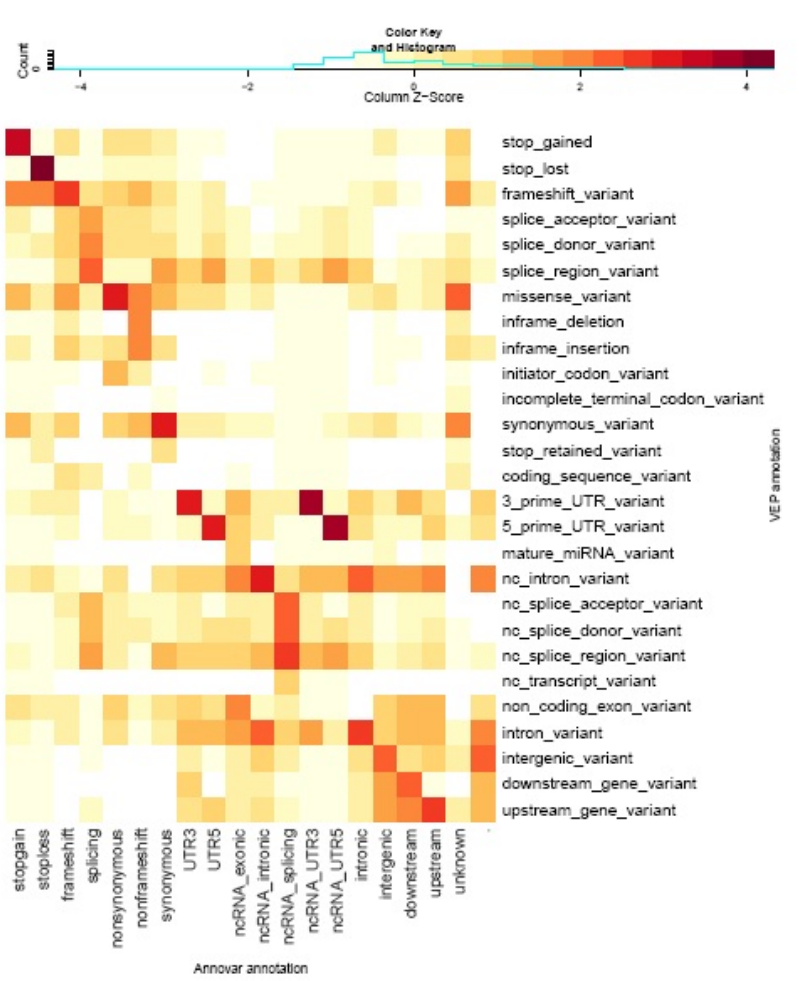
Nucleic Acids Research

Terms Used in Annovar and VEP

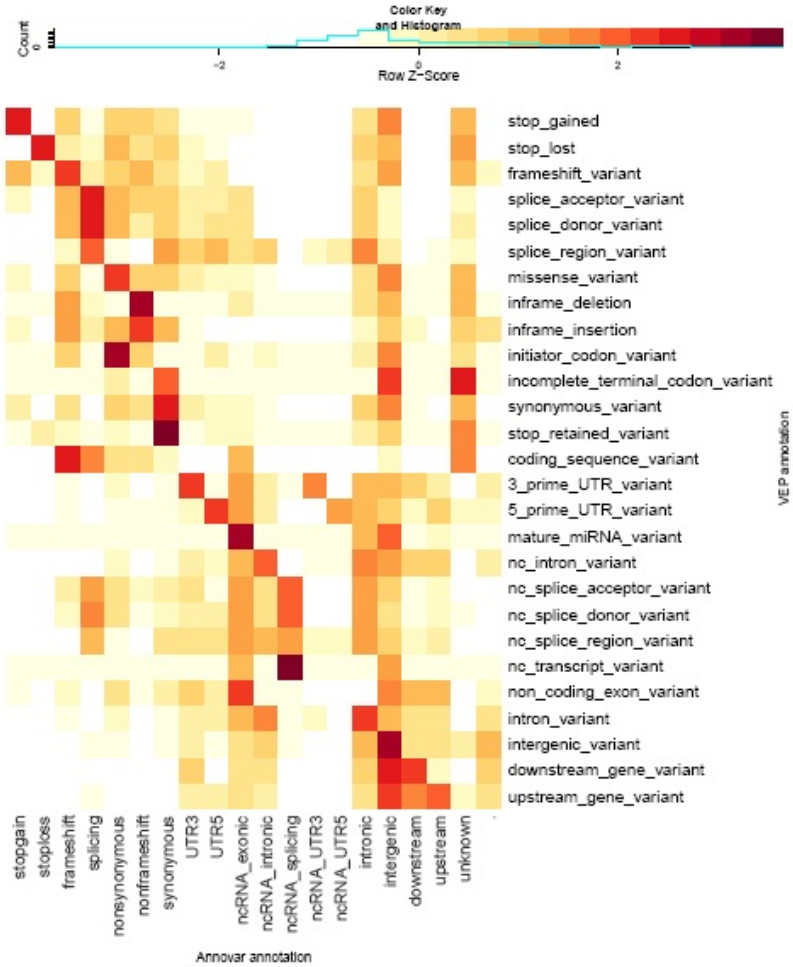
Category	ANNOVAR Terms	VEP Terms
Loss-of-function	"frameshift_deletion" "frameshift_insertion" "splicing" "stopgain_SNV" "stoploss_SNV"	"frameshift_variant" "splice_donor_variant" "splice_acceptor_variant" "stop_gained" "stop_lost" "transcript_ablation"
Missense/Nonsynonymous	"nonframeshift_deletion" "nonframeshift_insertion" "nonsynonymous_SNV"	"inframe_insertion" "inframe_deletion" "splice_region_variant" "initiator_codon_variant"
Synonymous	"synonymous_SNV"	"synonymous_variant" "stop_retained_variant"
Exonic	All of the above	All of the above plus "coding_sequence_variant" "incomplete_terminal_codon_variant"

Davis J McCarthy, *Genome Medicine* 2014. doi:10.1186/gm543

Heatmap of Annovar and VEP



Annovar-normalized

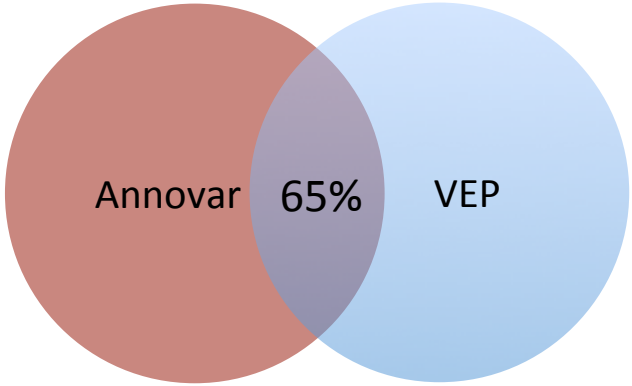


VEP-normalized

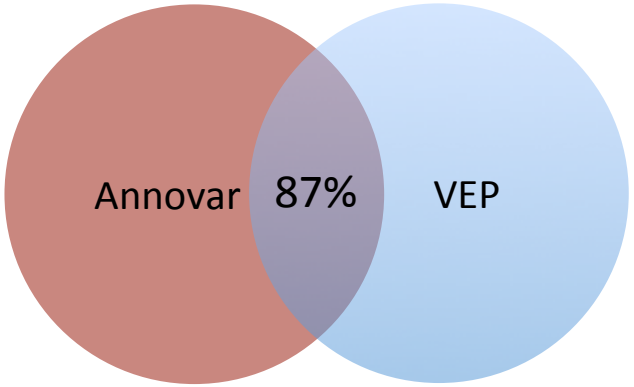
Davis J McCarthy, *Genome Medicine* 2014. doi:10.1186/gm543

Venn Diagram between Annovar and VEP using Ensembl

LOF variants



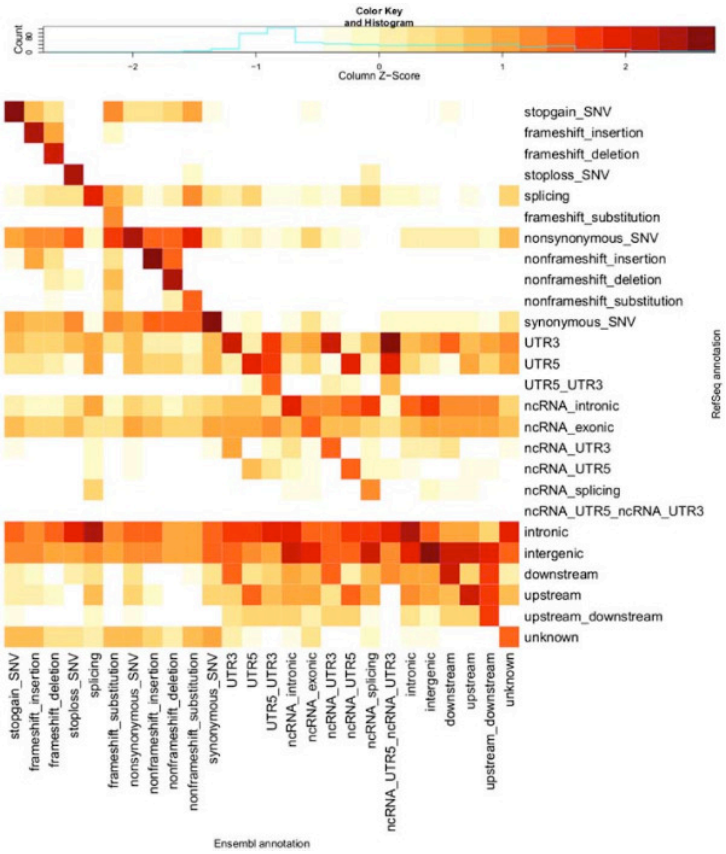
Exonic variants



Comparison between RefSeq and Ensembl in Annovar

RefSeq-normalized

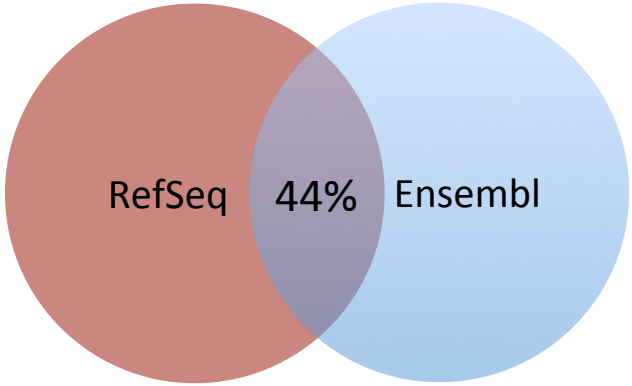
Ensembl-normalized



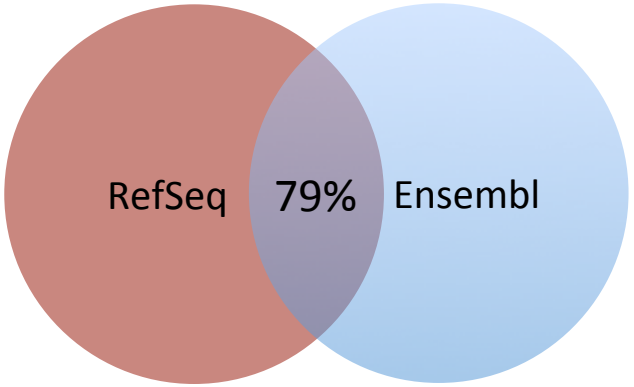
Davis J McCarthy, *Genome Medicine* 2014. doi:10.1186/gm543

Venn Diagram between RefSeq and Ensembl in Annovar

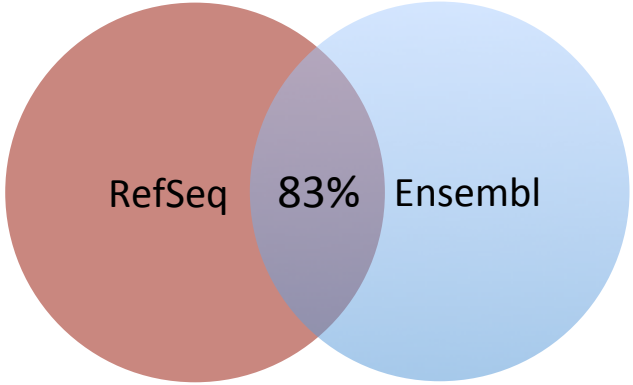
LOF variants



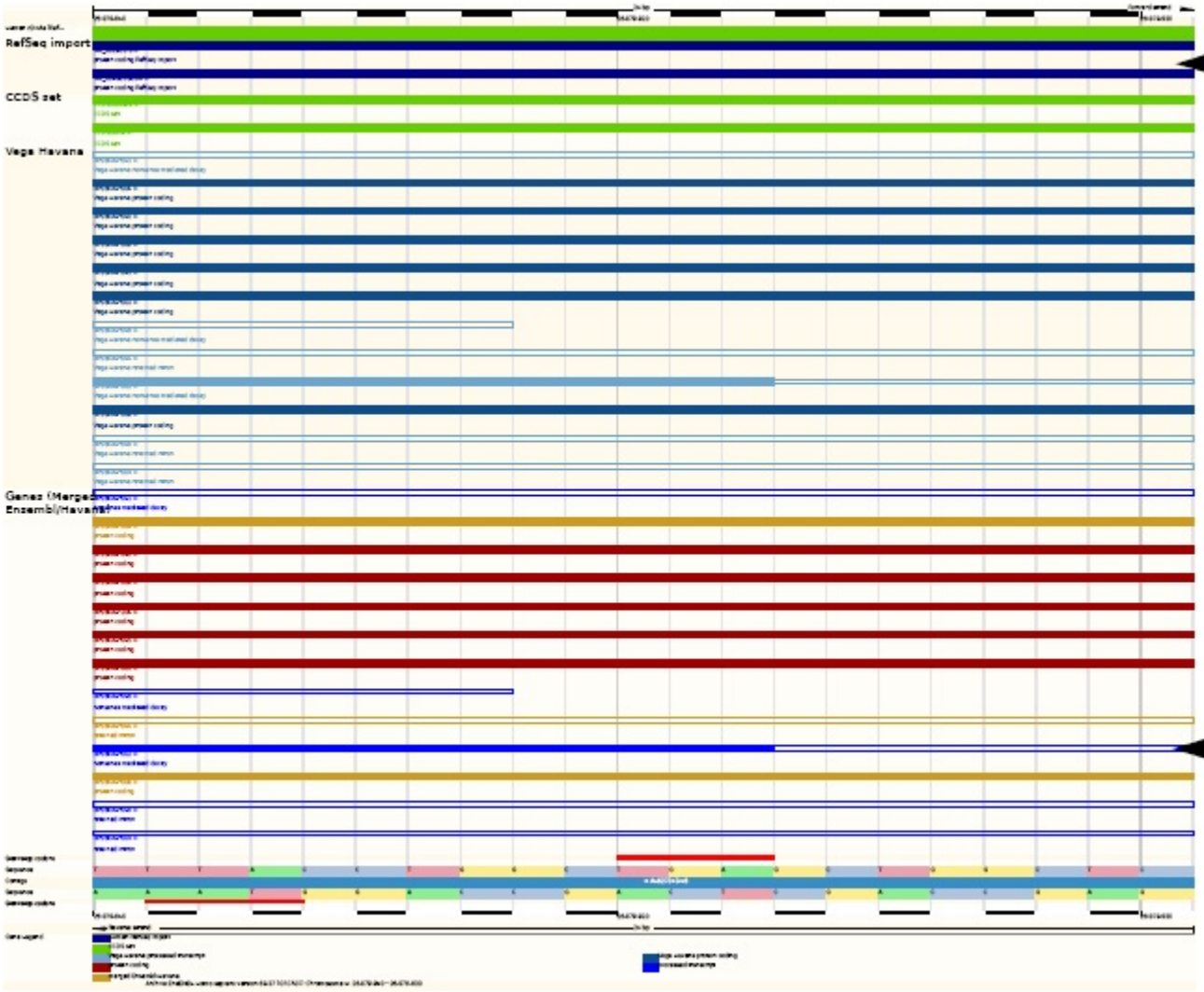
LOF + Nonsynonymous variants



Exonic variants



How different it could be between RefSeq and Ensembl

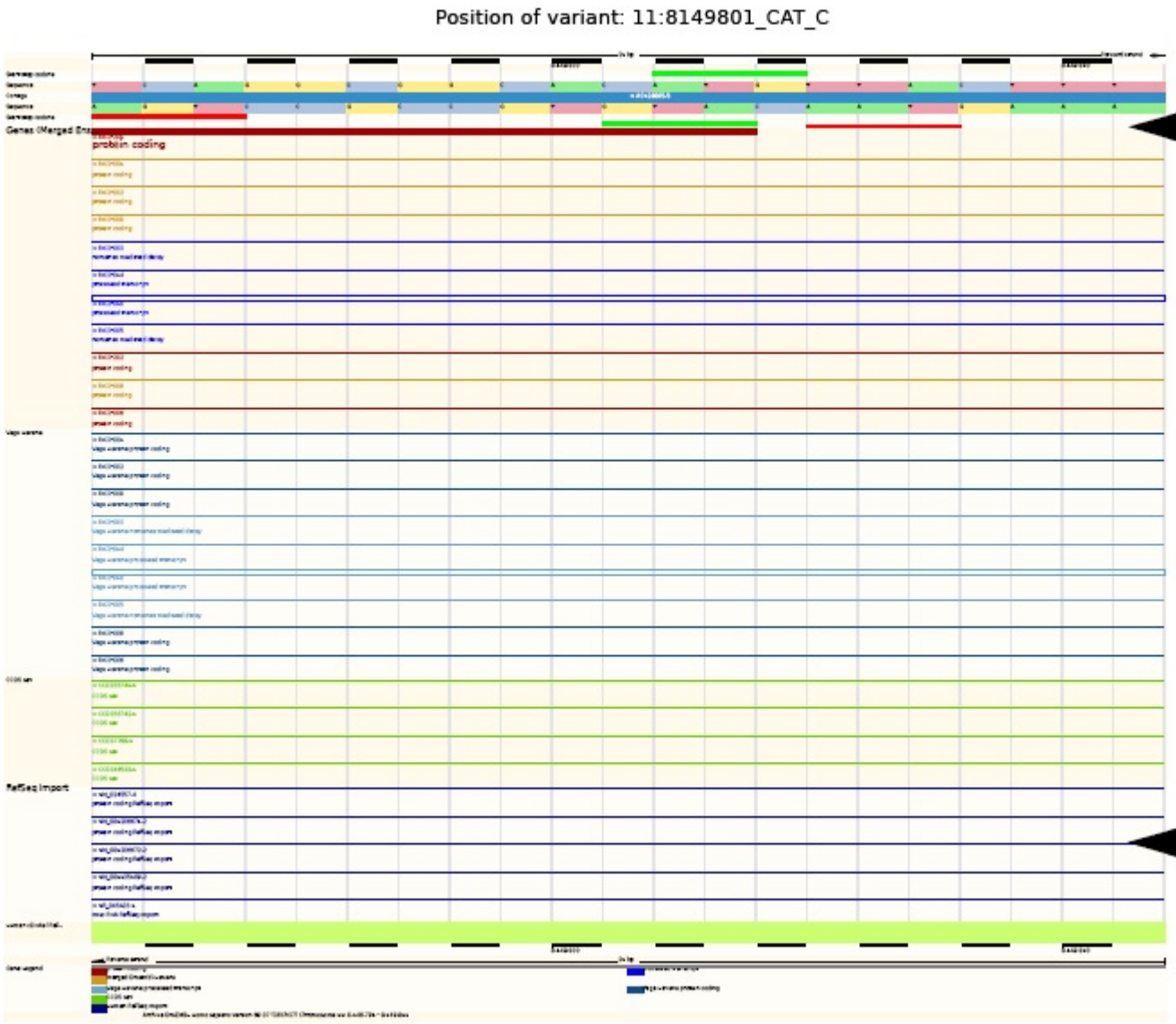


Transcripts used for RefSeq-based annotation "synonymous_SNV"

Transcript used for Ensembl-based annotation ENST00000374163 "stoploss_SNV"

Davis J McCarthy, *Genome Medicine* 2014. doi:10.1186/gm543

Another Example

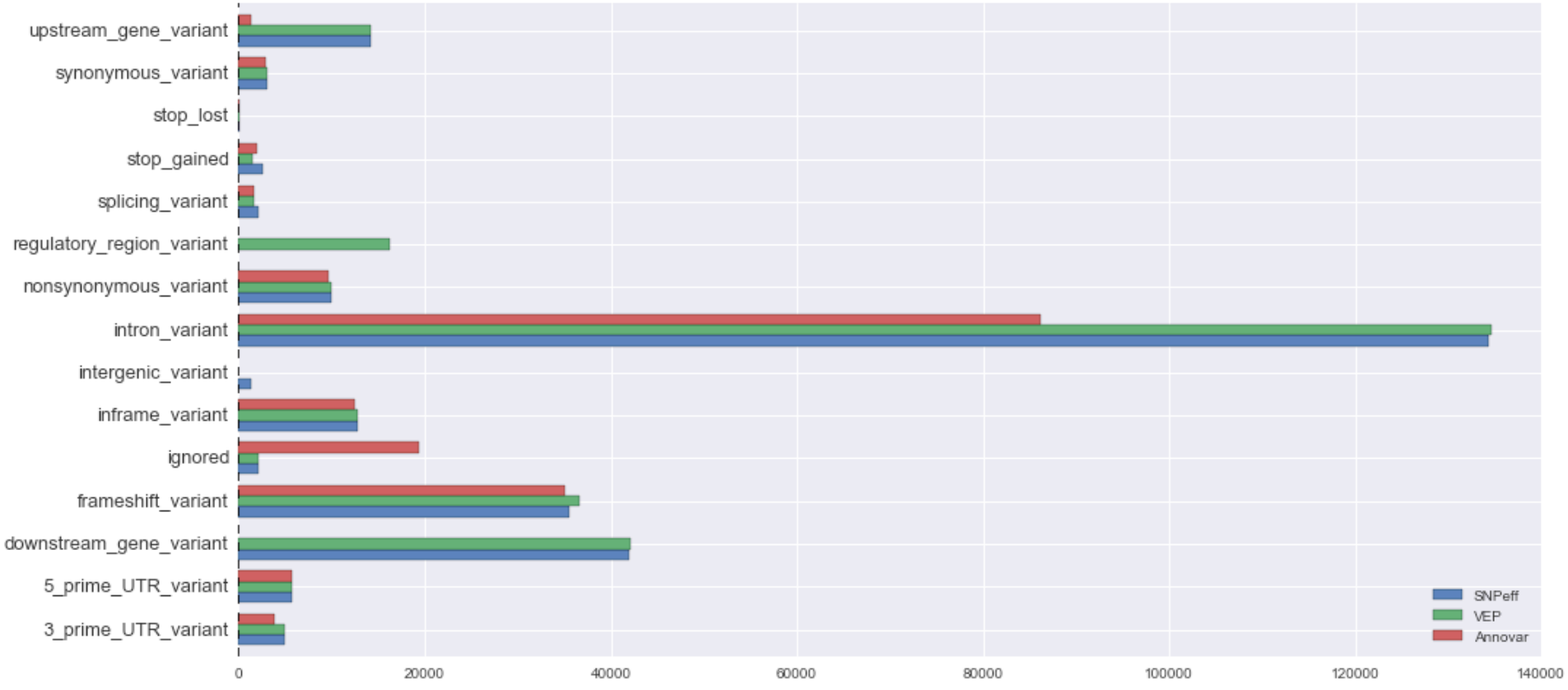


Davis J McCarthy, *Genome Medicine* 2014. doi:10.1186/gm543

Summary of SO Names in 3 Annotations

Normalized SO Name	ANNOVAR	VEP	SNPeff
frameshift_variant	frameshift_deletion, frameshift_insertion, frameshift_block_substitution	frameshift_variant	FRAME_SHIFT
stop_gained	stopgain	stop_gained	STOP_GAINED
stop_lost	stoploss	stop_lost	STOP_LOST
splicing_variant	splicing	splice_donor_variant, splice_acceptor_variant	SPLICE_SITE_DONOR, SPLICE_SITE_ACCEPTOR
inframe_variant	nonframeshift_deletion, nonframeshift_insertion	inframe_insertion, inframe_deletion	CODON_INSERTION, CODON_CHANGE_PLUS_CODON_INSERTION, CODON_DELETION, CODON_CHANGE_PLUS_CODON_DELETION
nonsynonymous_variant	nonsynonymous_SNV, nonframeshift_block_substitution	initiator_codon_variant, missense_variant, stop_retained_variant, incomplete_terminal_codon_variant	CODON_CHANGE, NON_SYNONYMOUS_CODING, NON_SYNONYMOUS_START, NON_SYNONYMOUS_STOP, START_LOST
synonymous_variant	synonymous_SNV	synonymous_variant	SYNONYMOUS_CODING, SYNONYMOUS_START, SYNONYMOUS_STOP
3_prime_UTR_variant	UTR3	3_prime_UTR_variant	UTR_3_PRIME, UTR_3_DELETED
5_prime_UTR_variant	UTR5	5_prime_UTR_variant	UTR_5_PRIME, UTR_5_DELETED, START_GAINED
upstream_gene_variant	upstream	upstream_gene_variant	UPSTREAM
downstream_gene_variant	downstream	downstream_gene_variant	DOWNSTREAM
regulatory_region_variant	N/A	regulatory_region_variant, regulatory_region_ablation, regulatory_region_amplification	REGULATION
intron_variant	intronic	intron_variant	INTRON, INTRON_CONSERVED
intergenic_variant	intergenic	intergenic_variant	INTERGENIC, INTERGENIC_CONSERVED
Ignored	unknown, exonic, ncRNA	transcript_ablation , coding_sequence_variant, splice_region_variant, feature_truncation, feature_elongation, TF_binding_site_variant, TFBS_amplification, TFBS_ablation, NMD_transcript_variant, nc_transcript_variant, non_coding_exon_variant, mature_miRNA_variant	EXON, GENE, EXON_DELETED, CDS, CHROMOSOME, SPLICE_SITE_REGION, SPLICE_SITE_BRANCH, SPLICE_SITE_BRANCH_U12, MICRO_RNA, INTRAGENIC

The concordance between the three algorithms in coding regions

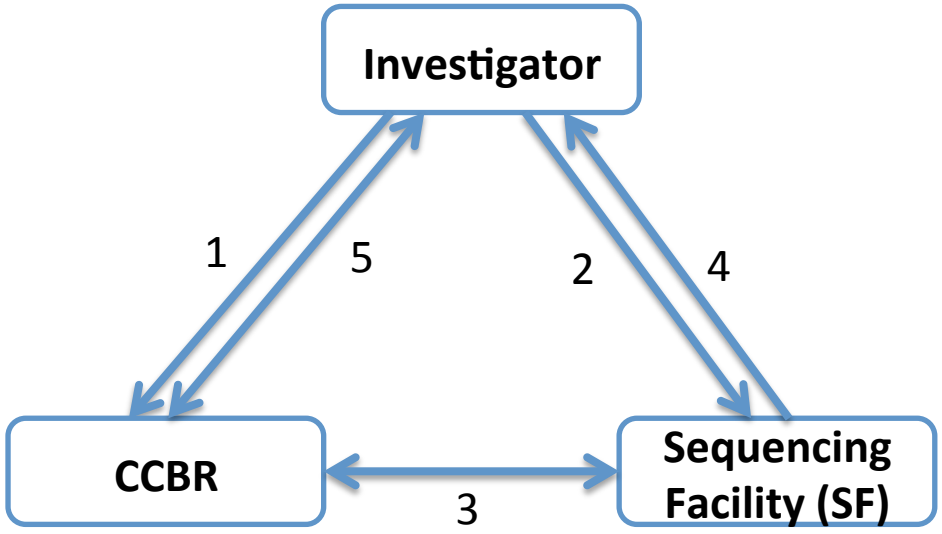


Andrew Jesaitis, http://blog.goldenhelix.com/?p=2486

Outlines

- Lesson learned from the bad experimental design
- Best practice in CCBR
- Functional annotation
- How to collaborate with CCBR – guide to success

What You Need to Do



- Visit us before you begin your experiment to let us know your goals and participation
- Submit your project request
  - https://bioinformatics.cancer.gov/project_submission
  - Submit your sequence to SF
- CCBR will start the analysis after receiving all datasets from SF
- CCBR and investigators will keep the close communication during the analysis

Analysis on Your Own

- Learn appropriate QC methods, and experimental designs

- Know what is in your tool box
  - Command line
    - Biowulf account
    - The working space under Biowulf
    - R/Bioconductor packages
  - GUI tools
    - Galaxy

- To take this further
  - Know how to run command line programs
  - Learn how to write scripts
  - Learn some different tools: Golden-Helix, GeneGrid, IVA, GRAVAT/MuPIT

Questions?

- CCBR home page:
  https://bioinformatics.cancer.gov/
- CCBR email:
  CCBR@mail.nih.gov
- Office location:
  - Building 37, room 3041
  - Building 41, room B620
- Office hours:
  Fridays 10:00am -12:00pm