Sequence Annotation: The basics

IBG 2015

Monkol Lek
20150304

BROAD INSTITUTE

Overview

- The overall goal of annotation
- Types of sequence annotations
- Common gene models used for annotations
- Challenges of sequence annotations

Overall goal of sequence annotation

Given a sequence variant what can we say about it's likely effect on gene function and phenotype?

Before sequence annotation:

11	104763117	rs497116	G	A	30784768.36	PASS	AC=16742;AN=17040

After sequence annotation:

11 104763117 rs497116 G A 30784768.36 PASS

AC=16742;AC_AFR=518;AC_AMR=182;AC_Adj=14187;AC_EAS=198;AC_FIN=1594;AC_Het=185;AC_Hom=7001;AC_NFE=4006;AC_OTH=184;AC_SAS=7505;AF=0.983;AN=17040;AN_AFR=604;AN_AMR=188;AN_Adj=14394;AN_EAS=198;AN_FIN=1594;AN_NFE=4008;AN_OTH=184;AN_SAS=7618;BaseQRankSum=1.68;ClippingRankSum=0.138;DB;DP=544013;FS=1.152;GQ_MEAN=218.13;GQ_STDDEV=192.36;Het_AFR=70;Het_AMR=6;Het_EAS=0;Het_FIN=0;Het_NFE=2;Het_OTH=0;Het_SAS=107;Hom_AFR=224;Hom_AMR=88;Hom_EAS=99;Hom_FIN=797;Hom_NFE=2002;Hom_OTH=92;Hom_SAS=3699;InbreedingCoeff=0.1753;MQ=59.71;MQ0=0;MQRankSum=0.294;NCC=78512;POSITIVE_TRAIN_SITE;QD=33.81;ReadPosRankSum=0.346;VQSLOD=6.03;culprit=QD;DP_HIST=745|578|78|35|44|59|107|208|325|485|595|675|676|688|637|559|372|269|199|1186,715|571|78|35|38|59|107|204|325|484|595|675|676|688|637|559|372|269|199|1186;GQ_HIST=23|437|292|353|166|41|43|24|12|16|11|9|23|15|10|26|22|19|35|6943,2|422|291|353|166|41|43|24|12|16|11|9|17|15|10|24|22|19|35|6940;CSQ=A|ENSG00000204403|ENST00000417998|Transcript|synonymous_variant&NMD_transcript_variant|724|468|156|C|tgC/tgT|rs497116&CM042005|1||-1|CASP12|HGNC|19004|nonsense_mediated_decay|||ENSP00000424963|CASPC_HUMAN||UPI0000228D0B|||4/8|||ENST00000417998.1:c.468C>T|ENST00000417998.1:c.468C>T(p.%3D)|G:0.0399|A:0.84|A:0.98|||||other||19200524&18852891&20210993&21706251&22615879&24307984||||||||

Types of annotations

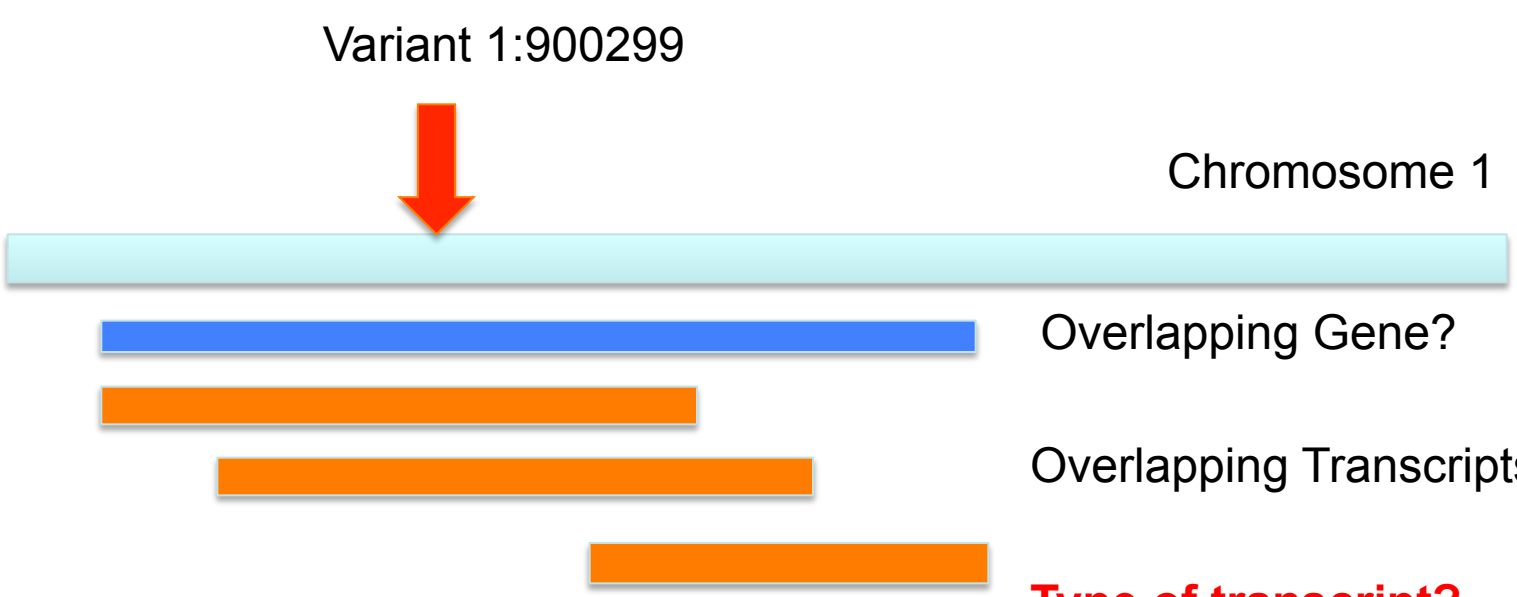
Basic annotation:

- Functional change

Pre-determined additional annotations:

- Allele frequency in the population (1KG, ESP, ExAC)
- Present in disease mutation databases (eg. ClinVar, HGMD)?
- Site conservation (eg. GERP, Phylop)
- Function impact predictions (eg. Polyphen, SIFT, CADD)

The process of functional annotations



In order to annotate need to know where the genes and transcripts are located

Major gene models for annotation:

- RefSeq
- Gencode
- CCDS

Others:

- UCSC
- ACE

Reference Sequence (RefSeq)

- Non-redundant
- Linked to nucleotide and protein sequences
- Manually curated – supported by cDNA, EST experimental data
- Prefix to distinguish between supported/curated (NM, NR and NP) vs predicted (XM, XR, XP)

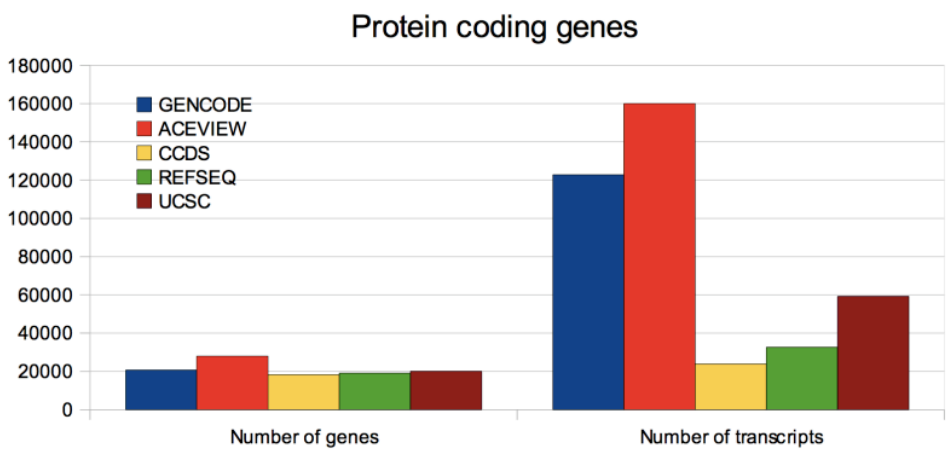
GENCODE (Encyclopædia of genes and gene variants)

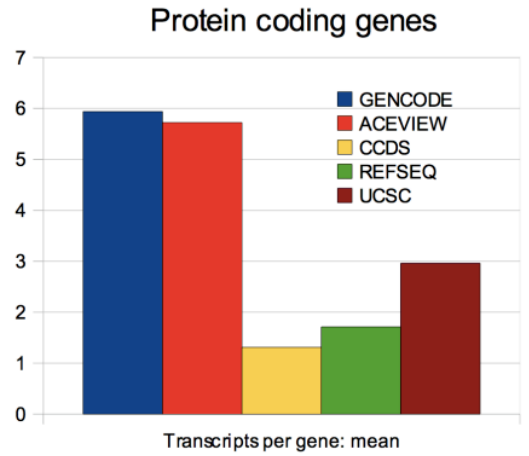
- Human reference gene set generated for the ENCODE project
- Merge of automatic (ENSEMBL) and extensive manual annotation (HAVANA)
- Experimental verification by RT-PCR and RACE of transcripts
- Manual annotation of all transcript biotypes (eg. protein-coding, processed-transcript, non-coding RNAs)
- Newer versions are for Build 38 only!

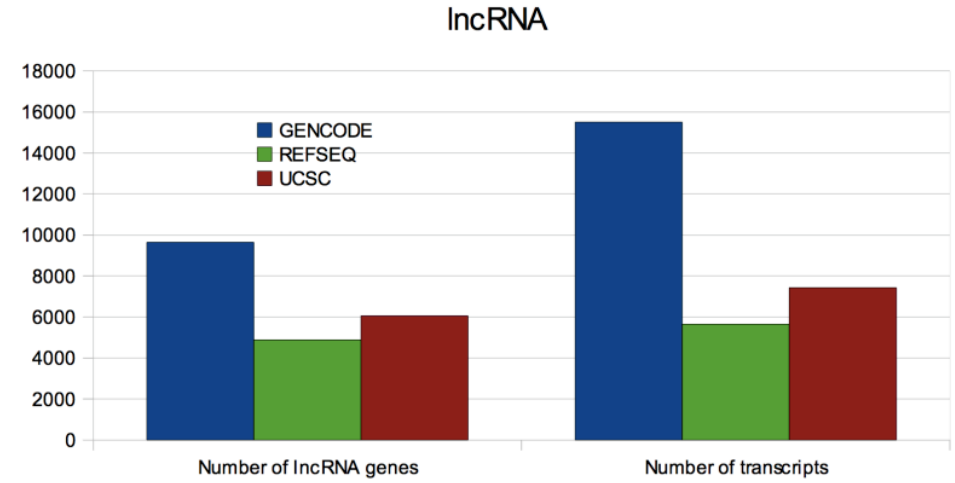
Consensus coding sequence (CCDS)

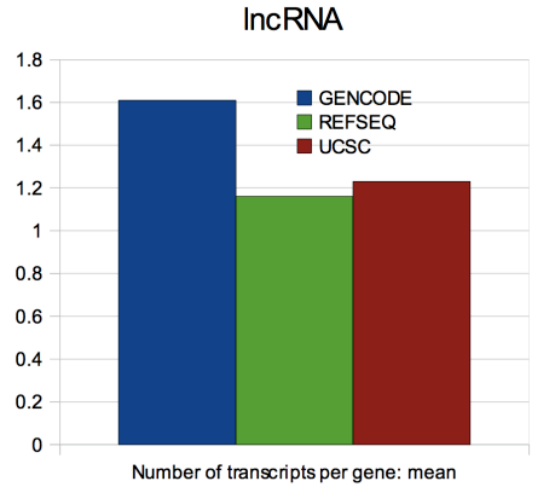
- Consensus between UCSC, Ensembl, Refseq and Havana
- Reference set must contain ATG, STOP on both Human and Mouse genomes
- High quality but few alternative splice variants an no UTRs, slow to increase in size

Summary metrics of Gene models

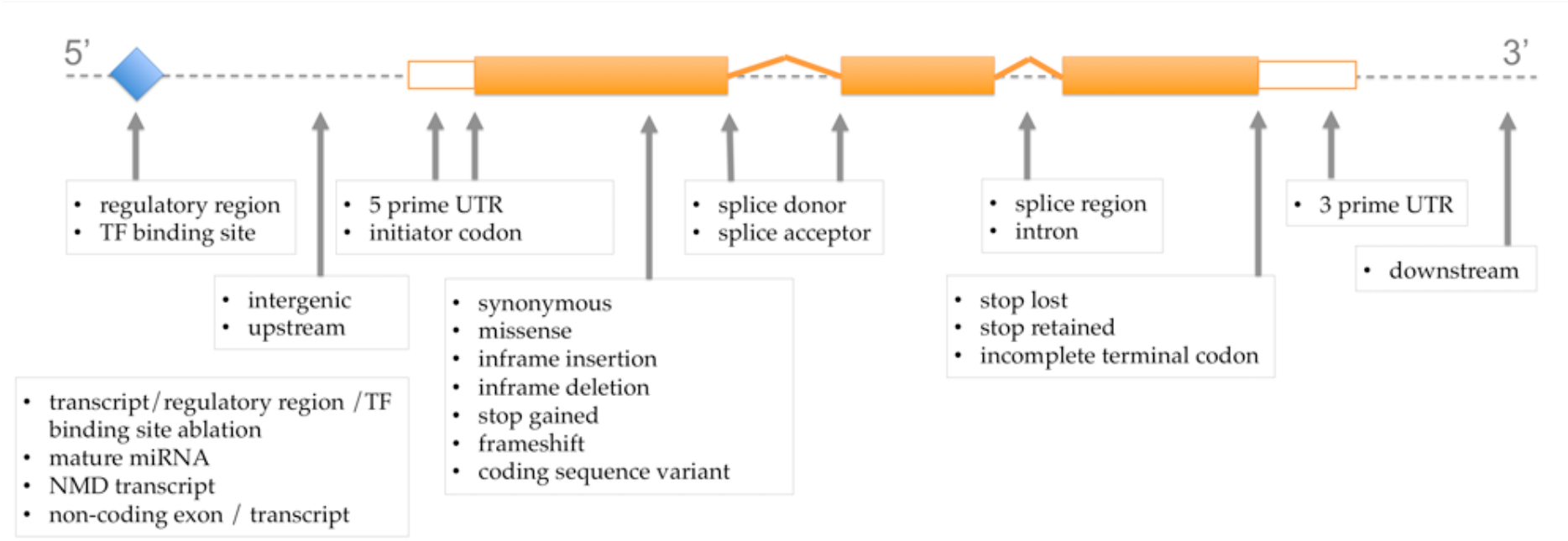








Functional annotation classes



http://useast.ensembl.org/info/docs/variation/predicted_data.html

Functional impact rank

* SO term	SO description
transcript_ablation	A feature ablation whereby the deleted region includes a transcript feature
splice_donor_variant	A splice variant that changes the 2 base region at the 5' end of an intron
splice_acceptor_variant	A splice variant that changes the 2 base region at the 3' end of an intron
stop_gained	A sequence variant whereby at least one base of a codon is changed, resulting in a premature stop codon, leading to a shortened transcript
frameshift_variant	A sequence variant which causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three
stop_lost	A sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript
initiator_codon_variant	A codon variant that changes at least one base of the first codon of a transcript
inframe_insertion	An inframe non synonymous variant that inserts bases into in the coding sequence
inframe_deletion	An inframe non synonymous variant that deletes bases from the coding sequence
missense_variant	A sequence variant, where the change may be longer than 3 bases, and at least one base of a codon is changed resulting in a codon that encodes for a different amino acid
transcript_amplification	A feature amplification of a region containing a transcript
splice_region_variant	A sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron
incomplete_terminal_codon_variant	A sequence variant where at least one base of the final codon of an incompletely annotated transcript is changed
synonymous_variant	A sequence variant where there is no resulting change to the encoded amino acid
stop_retained_variant	A sequence variant where at least one base in the terminator codon is changed, but the terminator remains
coding_sequence_variant	A sequence variant that changes the coding sequence
mature_miRNA_variant	A transcript variant located with the sequence of the mature miRNA
5_prime_UTR_variant	A UTR variant of the 5' UTR
3_prime_UTR_variant	A UTR variant of the 3' UTR
intron_variant	A transcript variant occurring within an intron
NMD_transcript_variant	A variant in a transcript that is the target of NMD
non_coding_exon_variant	A sequence variant that changes non-coding exon sequence
nc_transcript_variant	A transcript variant of a non coding RNA
upstream_gene_variant	A sequence variant located 5' of a gene
downstream_gene_variant	A sequence variant located 3' of a gene
TFBS_ablation	A feature ablation whereby the deleted region includes a transcription factor binding site
TFBS_amplification	A feature amplification of a region containing a transcription factor binding site
TF_binding_site_variant	In regulatory region annotated by Ensembl
regulatory_region_variant	A sequence variant located within a regulatory region
regulatory_region_ablation	A feature ablation whereby the deleted region includes a regulatory region

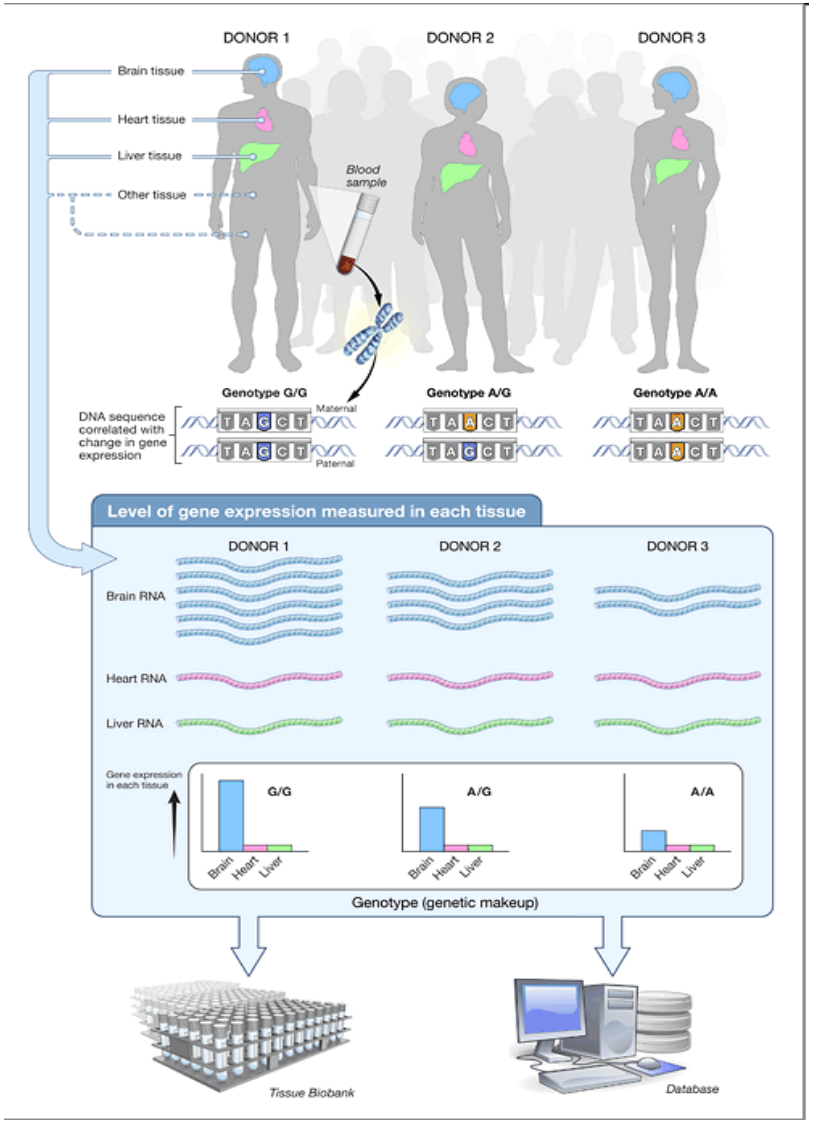
Challenge: Different gene model gives different annotation!

Table 1 Same software, different transcripts: RefSeq vs Ensembl by Annovar annotation category

	REF+ENS	REF	ENS	Match	REF match rate (%)	ENS match rate (%)	Overall match rate (%)
stopgain_SNV	15,835	14,183	14,960	13,308	93.83	88.96	84.04
frameshift_insertion	6,980	5,298	6,495	4,813	90.85	74.10	68.95
frameshift_deletion	7,491	4,547	7,380	4,436	97.56	60.11	59.22
stoploss_SNV	946	503	906	463	92.05	51.10	48.94
splicing	47,878	14,154	45,839	12,115	85.59	26.43	25.30
frameshift_substitution	1,960	195	1,947	182	93.33	9.35	9.29
nonsynonymous_SNV	321,669	291,898	315,592	285,821	97.92	90.57	88.86
nonframeshift_insertion	3,506	2,888	2,844	2,226	77.08	78.27	63.49
nonframeshift_deletion	5,136	3,321	4,963	3,148	94.79	63.43	61.29
nonframeshift_substitution	933	226	843	136	60.18	16.13	14.58
synonymous_SNV	178,559	167,561	172,463	161,465	96.36	93.62	90.43

McCarthy et. al. Genome Medicine 2014

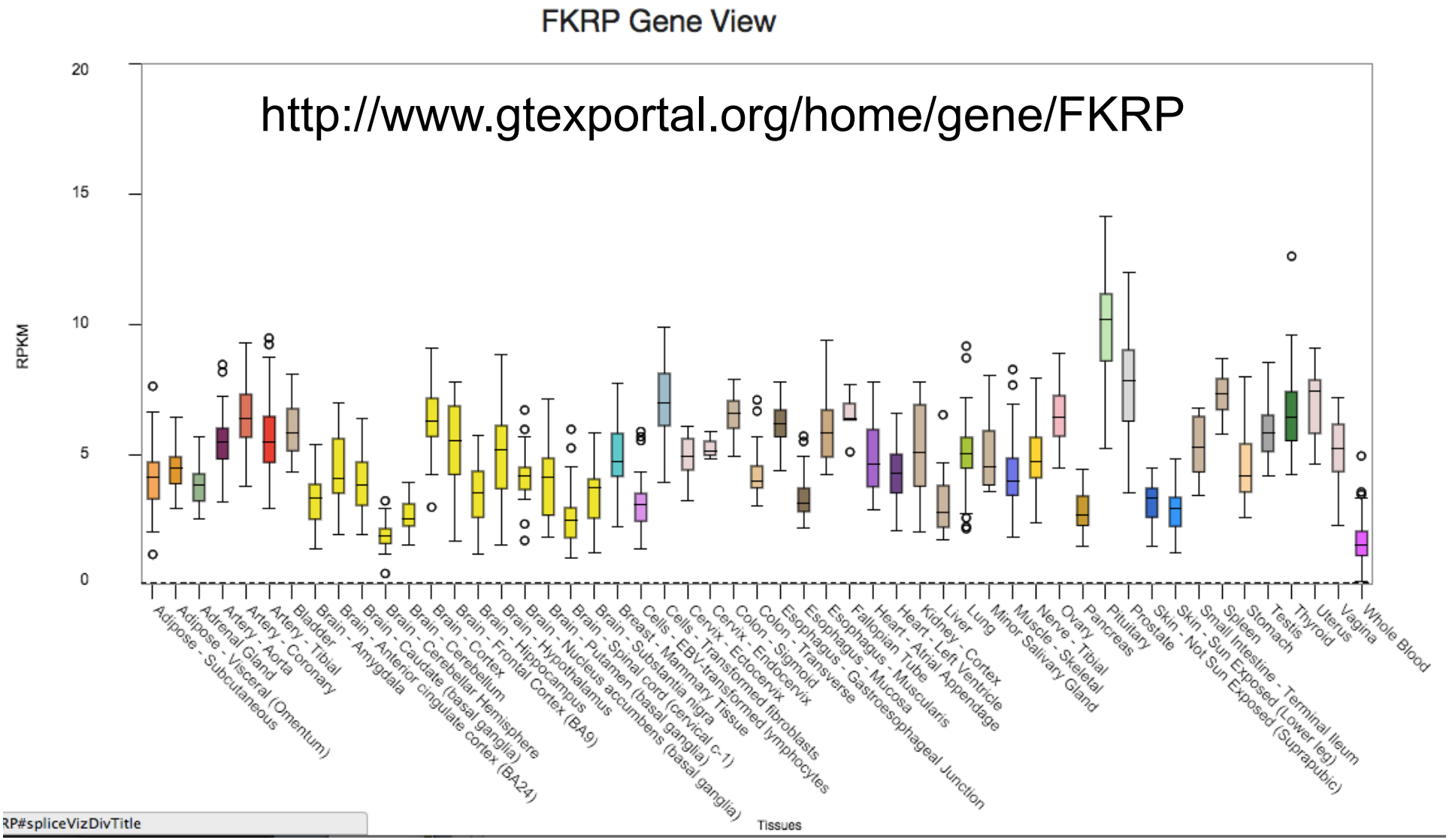
Challenge: Is the variant actually expressed in a relevant tissue?



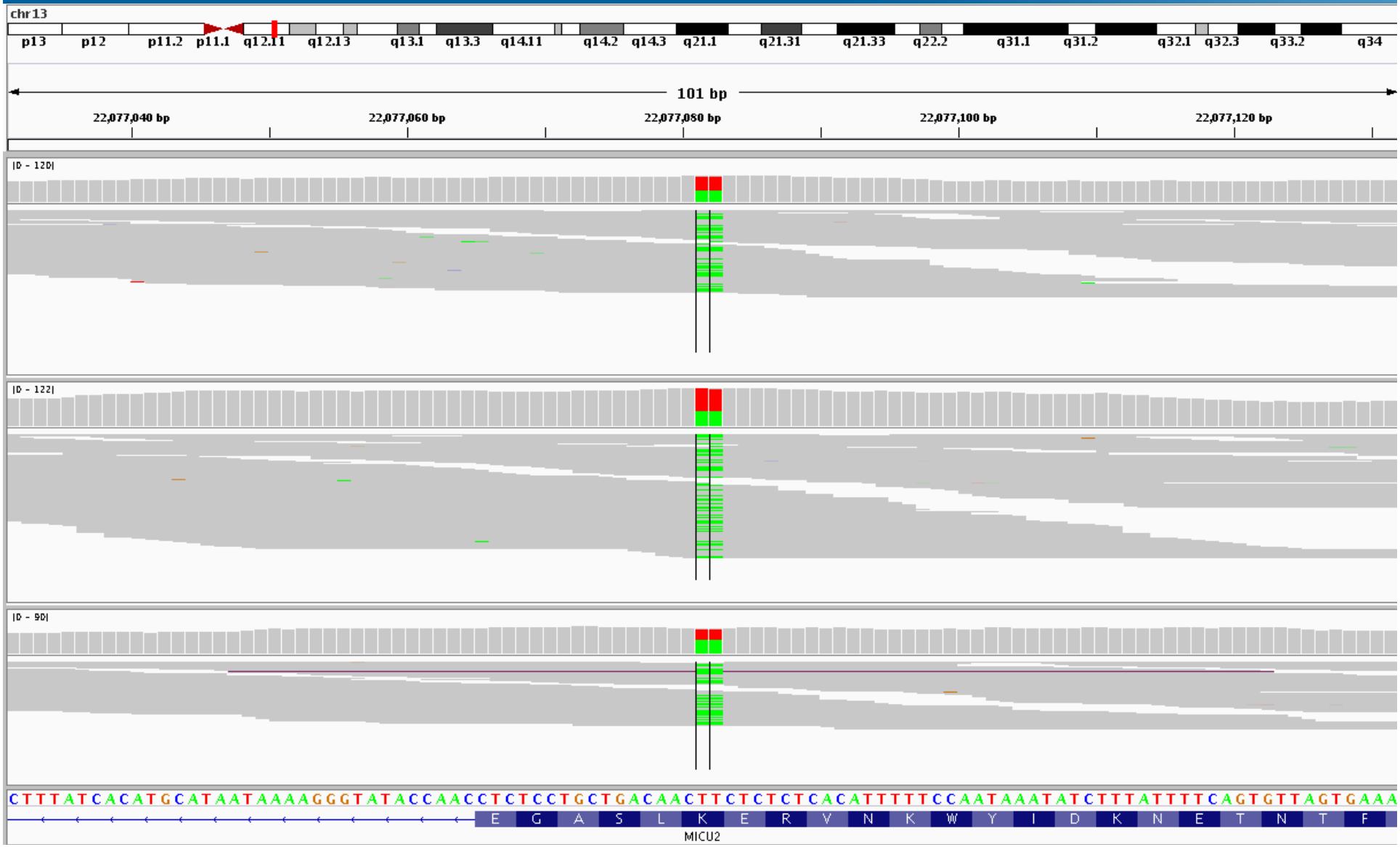
- Multiple tissues from autopsy and surgical donors
- Stringent protocol validation and quality control
- Deep RNA-seq on all tissues, plus 5M/ exomechip genotypes

	Completed: Through December 2012	Pilot Goal Feb '13	3 year Scale up goal
# Donors	176 (genotyped) (235 collected)	200	950
Arrays	955	1000	--
RNA Seq	1868	1868	24,759

Challenge: Is the variant actually expressed in a relevant tissue?



Challenge: Annotation tools assume that variants are independent



Challenge: Annotation tools struggle with multi-allelic variants

VCF line with deletion and SNP

1	1179747	.	CG	TG,C	979751.87	PASS	AC=476,1;

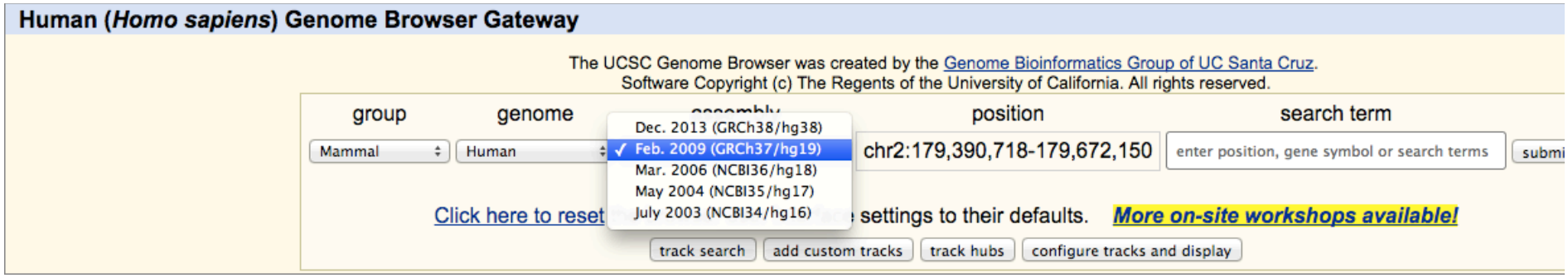
If the rare deletion did not exist in the cohort

1	1179747	.	C	T	979751.87	PASS	AC=476,1;

Cannot make an assumption that SNPs are on length 1!

And can not naively match on allele!

Challenge: A new human genome reference!



So which functional annotation do I use?

- If it is important not to miss potential LoF variants, then there are advantages to using ENSEMBL transcripts.

- If it is important to reduce false positives, then a carefully curated set of transcripts tailored to the study at hand may be preferred.

McCarthy et. al. Genome Medicine 2014

Annotation Tools

- Tools written in Perl
  - ANNOVAR
  - Variant Effect Predictor (VEP)
- Tools written in Java
  - SNPEff
- Tools written in C/C++
  - VAT

Variant Effect Predictor

Good Points

- Free to use for academics
- Maintained by a team!
- Uses Ensembl API Framework
- Plugin Framework to add extra functionality
  - e.g. LOFTEE (https://github.com/konradjk/loftee)
- Framework to add extra annotations
- Used by large consortia: Geuvadis, UK10K and ExAC

Variant Effect Predictor (VEP)

Bad Points

- Very slow! – can be split and jobs run in parallel
- Very large download
- Difficult to install
- Too much information and difficult to parse

ANNOVAR

Good Points

- Free to use for academics
- Easy to install and get working
- Easy to work with the output
- Framework to add extra annotations

ANNOVAR

Bad Points

- Maintained by one person!
- No framework to add extra functionality
- Default reference is hg18!
- Taking in VCF input and producing VCF output was an after thought