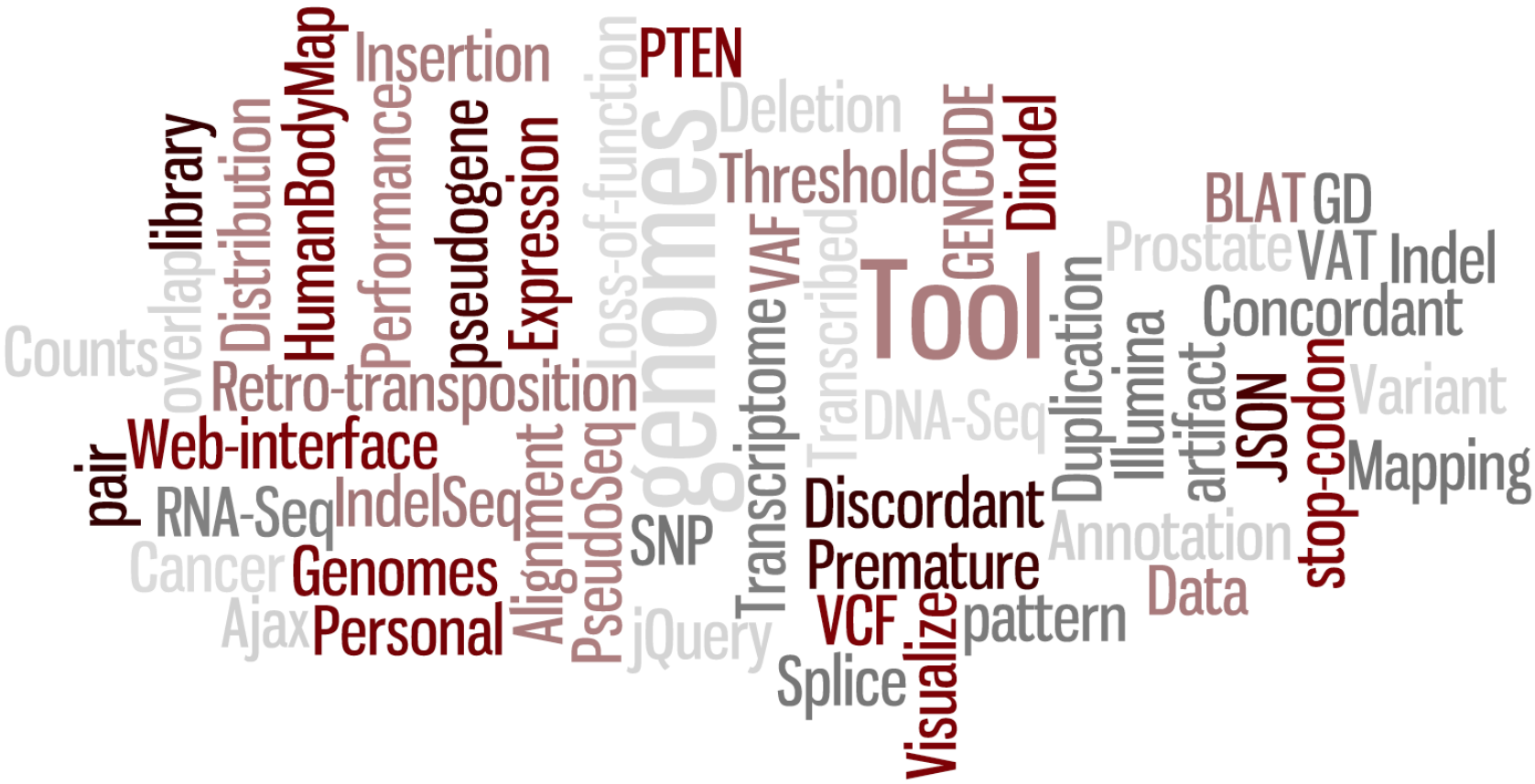


Lukas Habegger
Gerstein Lab
Group meeting
01/18/2011

Projects

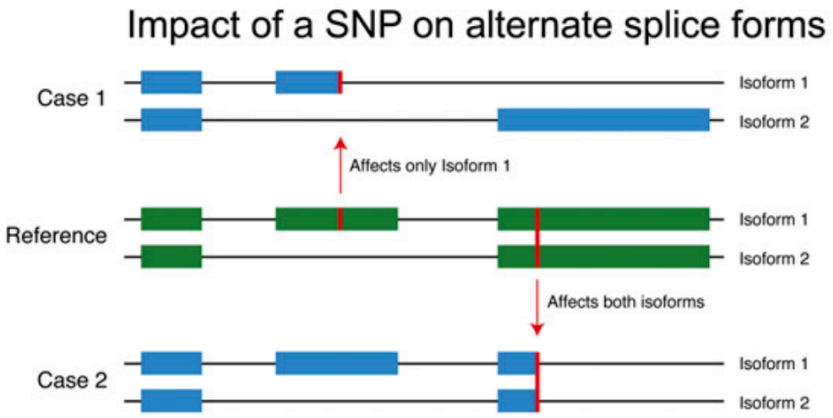
- Variation Annotation Tool (VAT)
- IndelSeq
- PseudoSeq

Objective

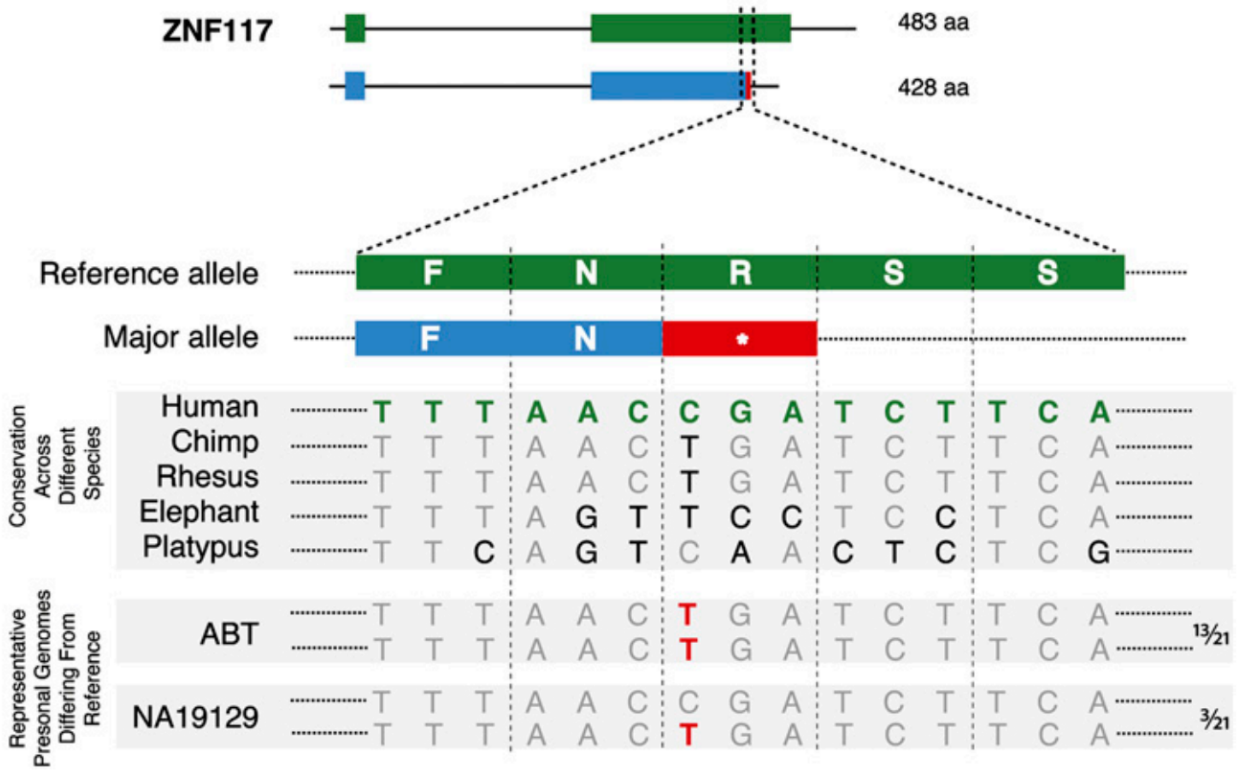
- To annotate genetic variants from personal genomes
  - SNPs
  - Indels
- Efficient algorithm
  - Command line
  - Web-interface
- Visualize the results

Types of variants

- SNPs
  - Synonymous
  - Non-synonymous
  - Premature stop
  - Removed stop
  - Splice overlap
- Indels
  - Insertions (frameshift, non-frameshift)
  - Deletions (frameshift, non-frameshift)
  - Splice overlap
  - Complex

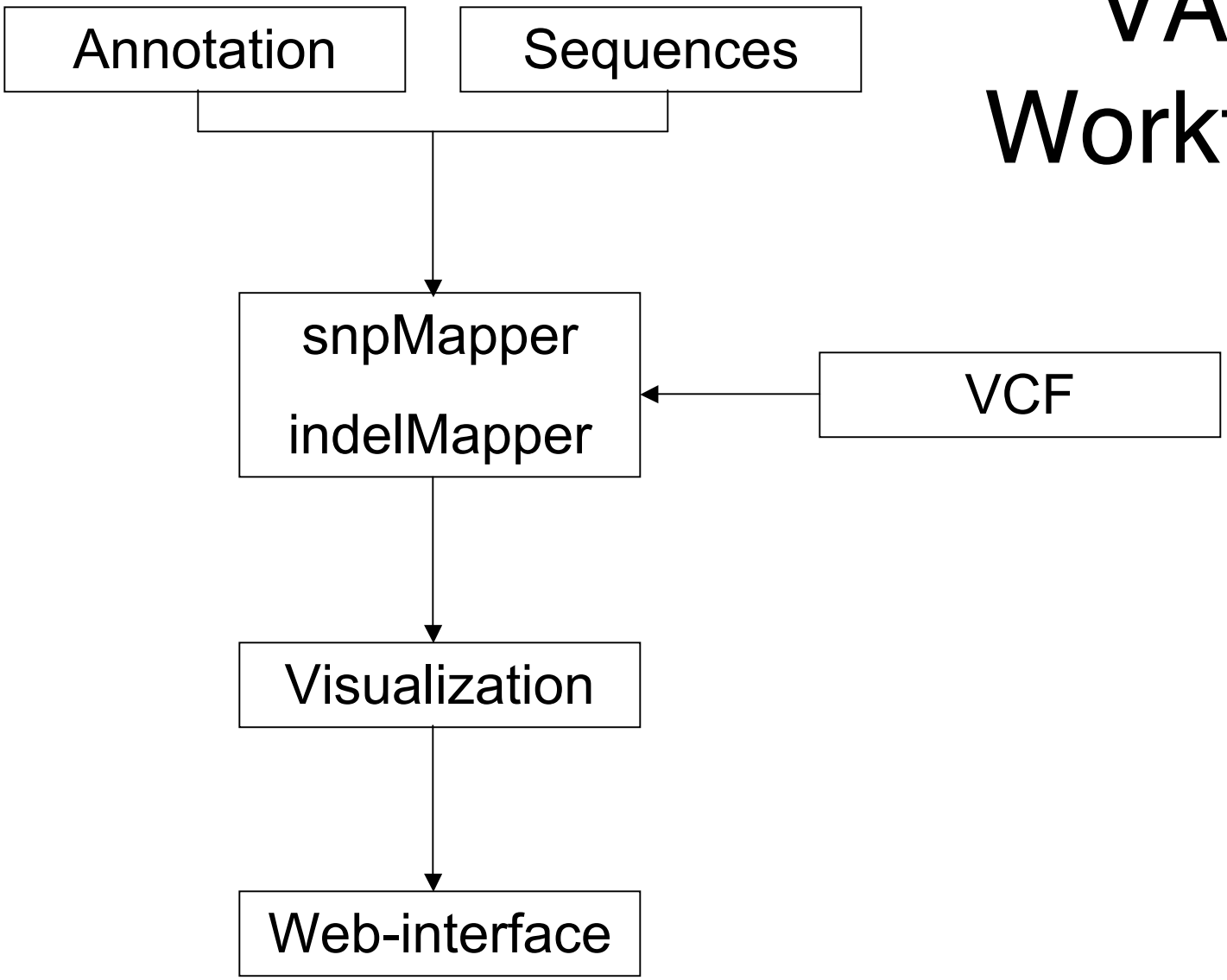


Manifestation of genetic variants



Balasubramanian et al., *Genes Dev.*, 2011

VAT Workflow



Implementation and Performance

- All programs are implemented in C
  - VCF (Variant call format) module
  - VAF (Variant annotation format) module

- Visualization
  - GD library

- Web-interface
  - jQuery, JSON, Ajax

- Performance (approximately **five times** faster than ANNOVAR)
  - Annotation of 10.5 SNPs (snpMapper)
    - Runtime: 100 seconds
    - Memory footprint: 150 MB

Web-interface

- Low-coverage
  - CEU
  - YRI
  - CHBJPT
- Trios
  - CEU
  - YRI

Results for data set: CEU.low_coverage

Show 25 entries Search:

Gene ID	Gene name	Number of transcripts	Number of synonymous SNPs	Number of nonsynonymous SNPs	Number of prematureStop SNPs	Number of removedStop SNPs	Number of spliceOverlaps	Number of insertions	Number of deletions	Details
ENSG00000000419	DPM1	6	0	1	0	0	0	0	0	Link
ENSG00000000457	SCYL3	4	5	2	0	0	0	0	0	Link
ENSG00000000460	C1orf112	4	2	3	0	0	0	0	0	Link
ENSG00000000938	FGR	5	1	0	0	0	0	0	0	Link
ENSG00000000971	CFH	5	4	5	0	0	0	0	0	Link
ENSG00000001036	FUCA2	5	1	5	0	0	0	0	0	Link
ENSG00000001084	GCLC	1	1	1	0	0	0	0	0	Link
ENSG00000001460	C1orf201	10	4	6	0	0	0	0	0	Link
ENSG00000001461	NIPAL3	9	1	1	0	0	0	0	0	Link
ENSG00000001561	ENPP4	2	1	1	0	0	0	0	0	Link
ENSG00000001626	CFTR	5	7	4	0	0	0	0	0	Link
ENSG00000001629	ANKIB1	2	1	1	0	0	0	0	0	Link
ENSG00000001630	CYP51A1	5	1	1	0	0	0	0	0	Link
ENSG00000001631	KRIT1	20	2	0	0	0	0	0	0	Link
ENSG00000002016	RAD52	4	0	1	0	0	0	1	0	Link
ENSG00000002330	BAD	3	1	1	0	0	0	0	0	Link
ENSG00000002726	ABP1	9	6	5	0	0	1	0	0	Link
ENSG00000002745	WNT16	3	0	3	0	0	0	2	0	Link
ENSG00000002746	HECW1	5	9	4	0	0	0	0	0	Link
ENSG00000002822	MAD1L1	15	5	2	0	0	0	0	0	Link
ENSG00000002834	LASP1	6	5	2	0	0	0	1	0	Link
ENSG00000002933	TMEM176A	5	2	6	0	0	0	0	0	Link
ENSG00000003056	M6PR	1	0	0	0	0	0	1	0	Link
ENSG00000003137	CYP26B1	2	1	1	0	0	0	0	0	Link
ENSG00000003147	ICA1	13	0	2	0	0	0	0	0	Link

Showing 1 to 25 of 15,853 entries

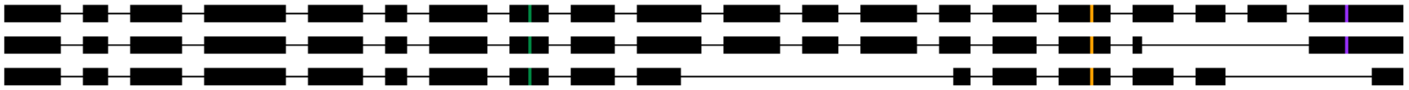
First Previous 1 2 3 4 5 Next Last

Results for ENSG00000068976 in data set CEU.low_coverage

Gene summary: PYGM (ENSG00000068976)

Transcript name	Transcript ID	Chromosome	Strand	Start	End	Number of exons	Transcript length
PYGM-001	ENST00000164139	chr11	-	64270709	64283946	20	2526
PYGM-002	ENST00000377432	chr11	-	64270709	64283946	18	2262
PYGM-201	ENST00000436572	chr11	-	64270709	64283946	16	1785

Graphical representation of variants



LEGEND FOR VARIATION TYPES:

spliceOverlap synonymous nonsynonymous prematureStop removedStop insertion deletion substitution

Detailed summary of variants

Chromosome	Position	Reference allele	Alternate allele	Identifier	Type	Fraction of transcripts affected	Transcripts	Details
chr11	64276502	G	C	.	synonymous	3/3	ENST00000164139 ENST00000377432 ENST00000436572	2526_1568_523_L->L 2262_1304_435_L->L 1785_827_276_L->L
chr11	64281910	C	A	.	nonsynonymous	3/3	ENST00000164139 ENST00000377432 ENST00000436572	2526_576_192_A->S 2262_312_104_A->S 1785_312_104_A->S
chr11	64283799	G	A	.	prematureStop	2/3	ENST00000164139 ENST00000377432	2526_147_49_R->* 2262_147_49_R->*

Next steps

- User-specified VCF files (file upload)
- Improvement of web-interface
  - Hyper-linking
- Population statistics for variants

Projects

- Variation Annotation Tool (VAT)
- IndelSeq
- PseudoSeq

Objective

- Analysis of indels in cancer genomes
- Manifestation of indels in the transcriptome:
  - Loss of function
    - Frameshifts
    - Disruption of splice sites
  - Trace indels: germline / somatic / transcriptome
  - Effect on gene expression
    - Does it affect all transcripts or just a subset?

Data

- Collaboration with Mark Rubin at Cornell
- 7 Matched prostate tissue samples
  - Normal (germline)
  - Tumor
- DNA-Seq done at the Broad
  - 30x coverage using Illumina
- RNA-Seq done at Cornell
  - Tumor
  - Normal (for a subset of samples)

Dindel

- **Inputs**:
  - BAM file (read alignments)
  - FASTA file of the reference genome

- **Workflow**:
  - **Stage 1**: Extract *candidate indels* from BAM file
  - **Stage 2**: Group *candidate indels* into windows
  - **Stage 3**: Construct *candidate haplotypes* based on indels; realign reads to *candidate haplotypes* (computationally intense)
  - **Stage 4**: Indel calling and filtering; output in VCF format

Number of genome-wide indel calls (after PASS filtering)

Sample	Normal	Tumor
STID0000000508	422473	448953
STID0000001701	458834	461414
STID0000001783	443530	435633
STID0000002832	450505	445596
STID0000003027	446443	439204
STID0000003043	453722	452375

Comparing indel calls: normal vs. tumor

Number of indel calls (genome-wide) after PASS filtering: all

	Unique Normal		Unique Tumor		Shared		
Sample	**Counts**	**Fraction**	**Counts**	**Fraction**	**Counts**	**Fraction**	**Total**
STID0000000508	41631	0.08	68099	0.14	380827	0.78	490557
STID0000001701	45520	0.09	48102	0.09	413261	0.82	506883
STID0000001783	60773	0.12	52907	0.11	382711	0.77	496391
STID0000002832	55158	0.11	50244	0.10	395326	0.79	500728
STID0000003027	51326	0.10	44076	0.09	395099	0.81	490501
STID0000003043	41823	0.08	40468	0.08	411872	0.83	494163

Number of indel calls (genome-wide) after PASS filtering: insertions only

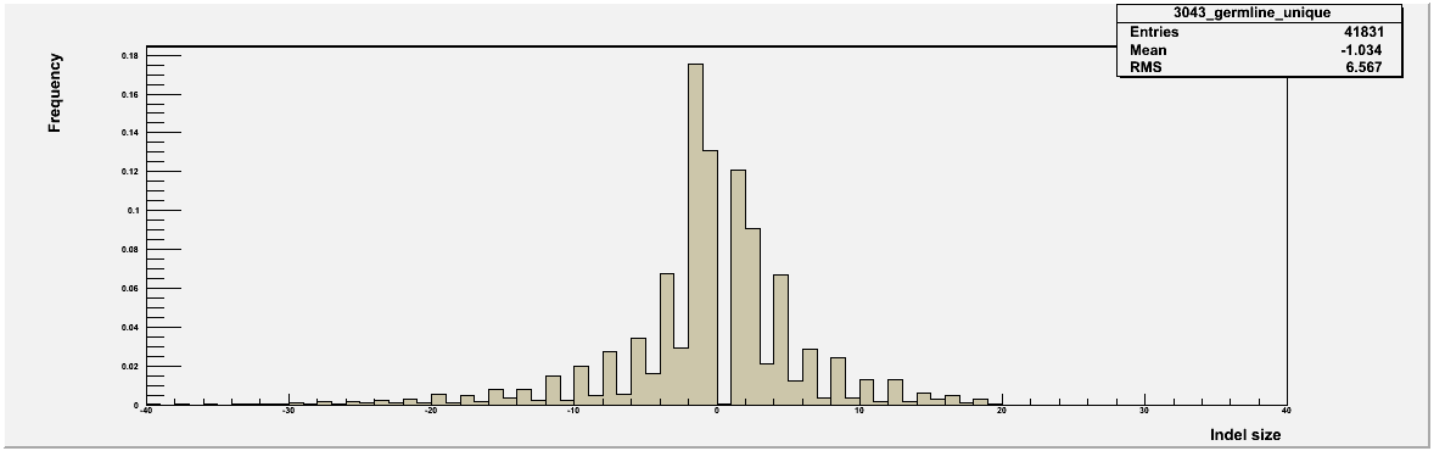
	Unique Normal		Unique Tumor		Shared		
Sample	**Counts**	**Fraction**	**Counts**	**Fraction**	**Counts**	**Fraction**	**Total**
STID0000000508	20416	0.09	25344	0.11	191128	0.81	236888
STID0000001701	19674	0.08	20943	0.09	201581	0.83	242198
STID0000001783	30691	0.13	27053	0.11	183948	0.76	241692
STID0000002832	28562	0.12	25934	0.11	189353	0.78	243849
STID0000003027	23449	0.10	20099	0.09	192577	0.82	236125
STID0000003043	17790	0.08	17090	0.07	201361	0.85	236241

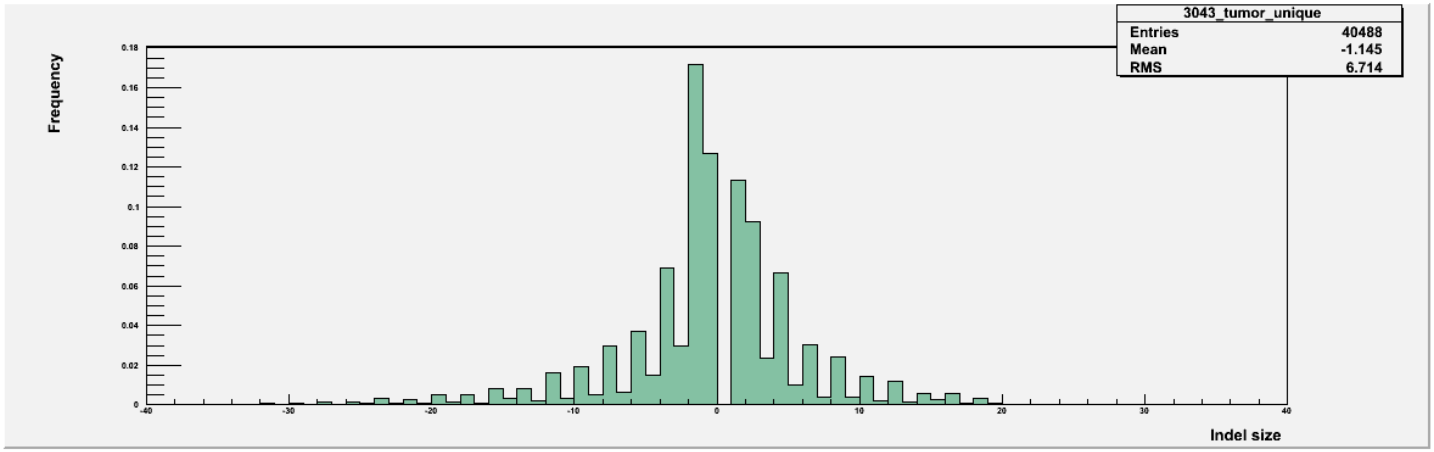
Number of indel calls (genome-wide) after PASS filtering: deletions only

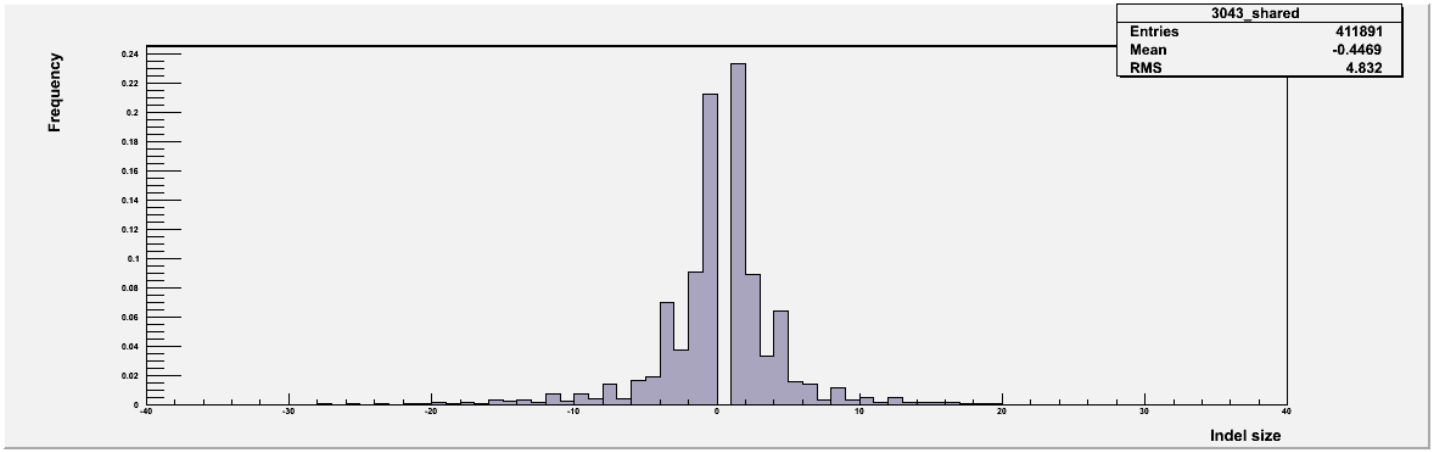
	Unique Normal		Unique Tumor		Shared		
Sample	**Counts**	**Fraction**	**Counts**	**Fraction**	**Counts**	**Fraction**	**Total**
STID0000000508	21848	0.09	43387	0.17	188998	0.74	254233
STID0000001701	26434	0.10	27731	0.10	211020	0.80	265185
STID0000001783	30500	0.12	26282	0.10	198282	0.78	255064
STID0000002832	27004	0.10	24706	0.10	205501	0.80	257211
STID0000003027	28383	0.11	24485	0.10	201938	0.79	254806
STID0000003043	24413	0.09	23767	0.09	210069	0.81	258249

Number of indel calls (genome-wide) after PASS filtering: 1nt indels only

	Unique Normal		Unique Tumor		Shared		
Sample	**Counts**	**Fraction**	**Counts**	**Fraction**	**Counts**	**Fraction**	**Total**
STID0000000508	11491	0.06	13604	0.07	177082	0.88	202177
STID0000001701	11260	0.05	11608	0.06	183374	0.89	206242
STID0000001783	18395	0.09	14783	0.07	172323	0.84	205501
STID0000002832	14779	0.07	13398	0.06	178140	0.86	206317
STID0000003027	15455	0.08	11442	0.06	176312	0.87	203209
STID0000003043	10710	0.05	9870	0.05	183218	0.90	203798

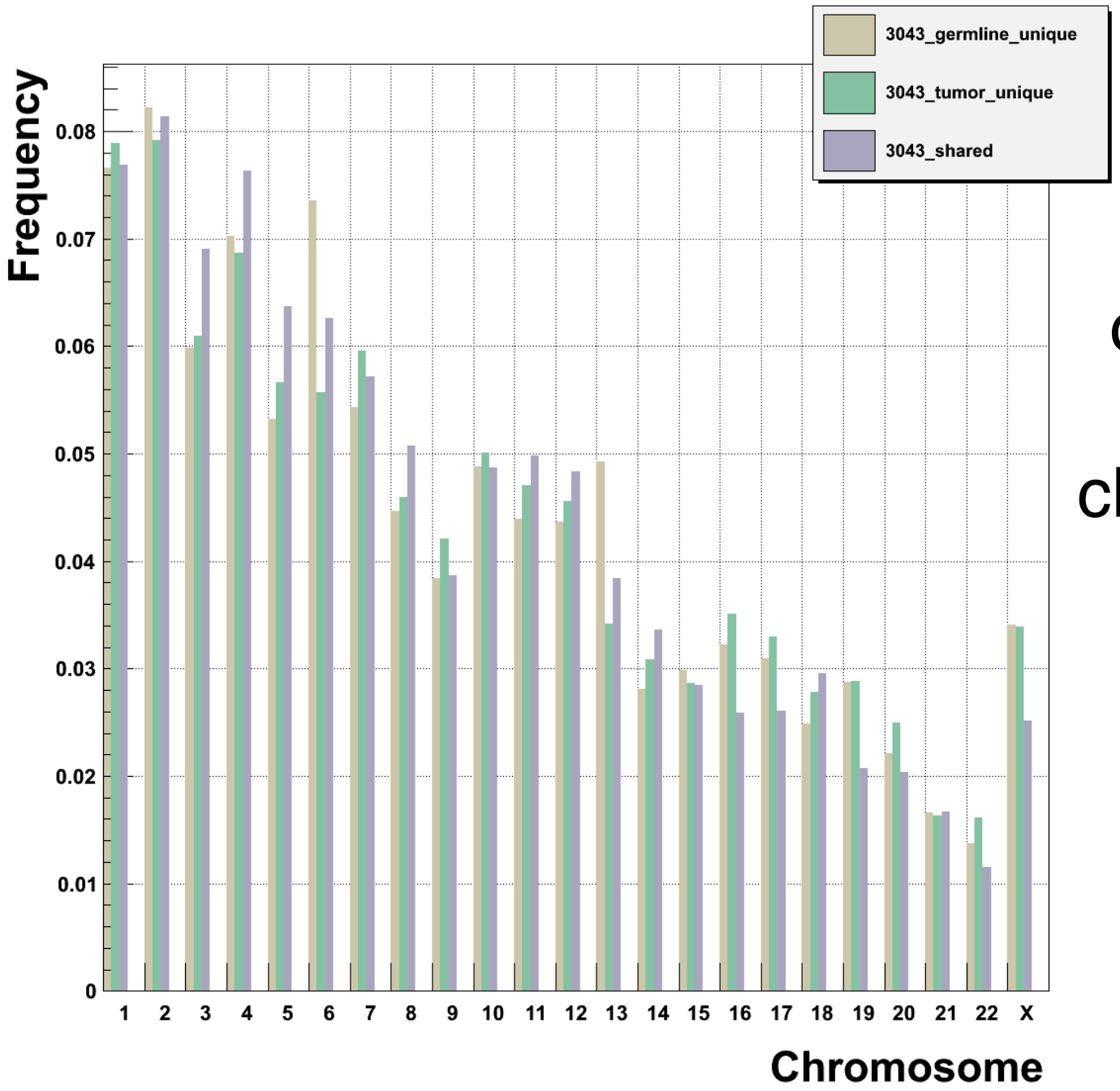




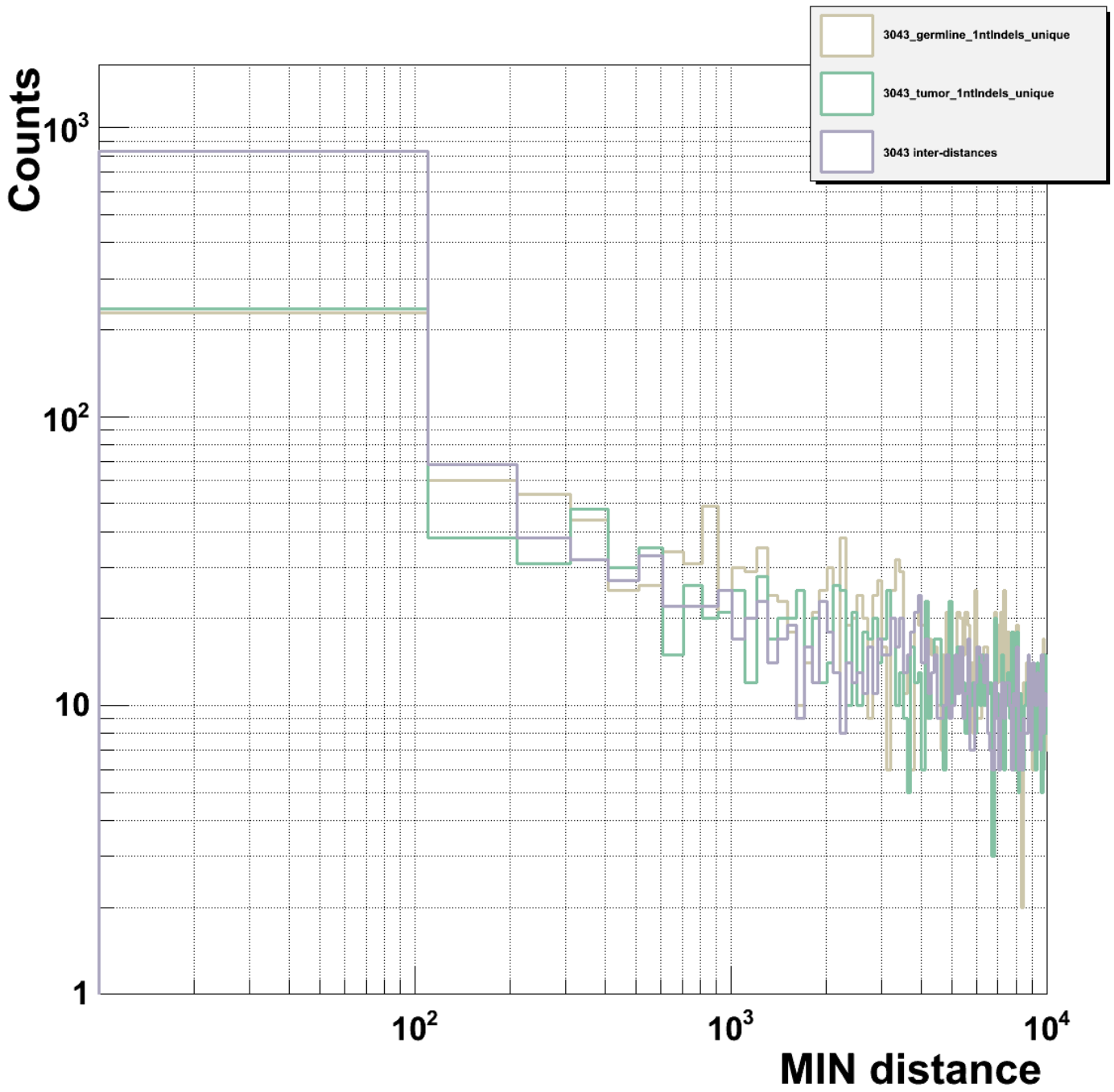


Indel size distribution

Periodicity of 2:
+/-2, +/- 4, +/- 6,…



Indel distribution by chromosome



Distribution of relative distances between indels

Number of indels in genes

- Annotation of indels using VAT
  - GENCODE v3b

Sample	Normal	Tumor
STID0000000508	695	781
STID0000001701	809	821
STID0000001783	781	792
STID0000002832	842	837
STID0000003027	772	752
STID0000003043	707	720

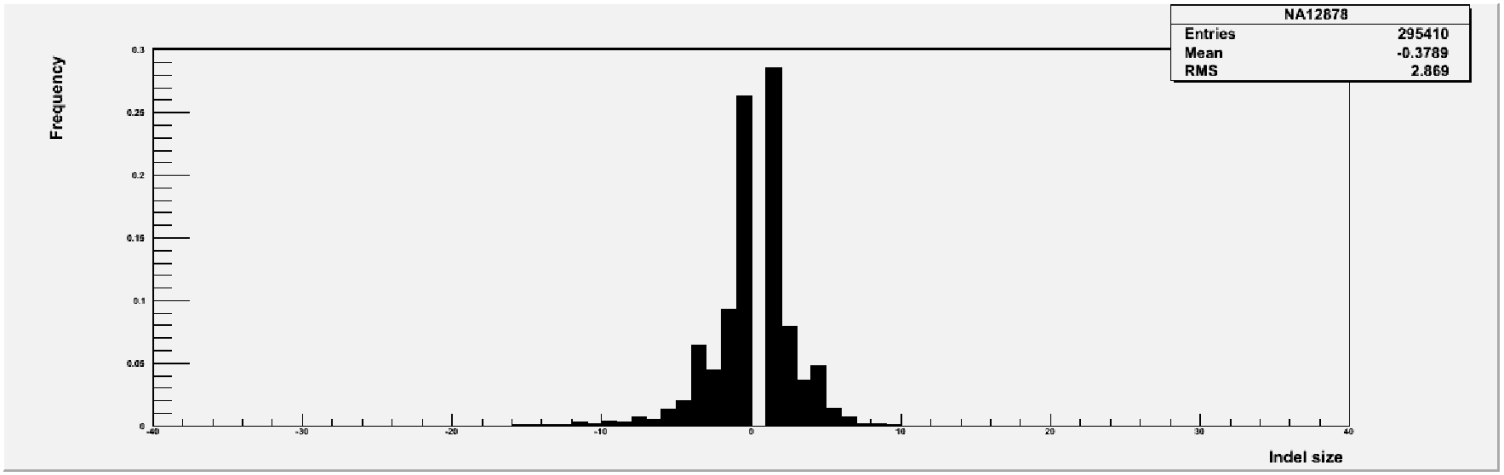
Comparison with NA12878

Indels form 1000 genomes CEU trio	
Number of indel calls in TRIO after PASS filtering	411611
Number of indel calls in TRIO after PASS filtering where NA12878 has a 0/0 genotype	83084
Number of indel calls in NA12878 after PASS filtering	**328527**

Indels processed at Yale	
Number of indel calls in NA12878 after PASS filtering	**297240**

Unique indels form 1000 genomes CEU trio	**Unique indels processed at Yale**	**Shared**
140568	26187	271043

91% of Yale indels are shared between the two sets



Affimetrix SNP data

Sample	SNPs shared
STID0000000508	0.997
STID0000001701	0.996
STID0000001783	0.990
STID0000002832	0.997
STID0000003027	0.983
STID0000003043	0.998

- The fraction of shared variants is significantly lower for indels compared to SNPs

Next steps

- Analyze the effect on gene expression
  - Map RNA-Seq reads to indel haplotypes
- Investigate indels that are unique to the tumor

Projects

- Variation Annotation Tool (VAT)
- IndelSeq
- PseudoSeq

Introduction

- Pseudogenes are created by
  - Duplication
  - Retro-transposition
- Characterized by the accumulation of mutations such as premature stop codons
- Most pseudogenes are presumed dead
- Some have acquired new functions and play important regulatory roles
  - Transcribed pseudogenes (born-again)
  - Example: PTENP1

Objective

- Identification of transcribed pseudogenes
- Compare the expression patterns of the pseudogene to its corresponding parent gene
  - Use of multiple RNA-Seq samples
  - Distinguish between potentially transcribed pseudogenes and mapping artifacts

Workflow

1. Obtain pgenes from GENCODE (v3b)

2. Exclude pgenes that overlap with coding sequences

3. Extract pseudo-transcript sequences and align them against the human reference using BLAT

4. Obtain all alignment pairs (pgene and hit)
    - Must have at least one alignment block that is larger than 100 nucleotides
    - **Note: A pgene can produce multiple alignment pairs**

Workflow continued

5. For each alignment pair, extract the RNA-Seq signal (signal track of mapped reads) for both the pseudogene and the corresponding hit

6. Decorate alignment pairs, calculate and include additional information:
   - Calculate statistics for the RNA-Seq signal (pgene and hit)
   - Calculate correlations
   - Intersect with annotation sets

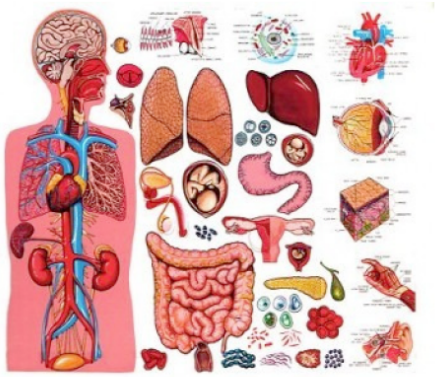
7. Calculate a score for each alignment pair

8. Generate an image for each alignment pair

9. Web-interface

Human Body Map

Human Body Map 2.0 Project



Tissues:

- Adrenal
- Adipose
- Brain
- Breast
- Colon
- Heart
- Kidney
- Liver
- Lung
- Lymph Node
- Ovary
- Prostate
- Skeletal Muscle
- Testis
- Thyroid
- White Blood Cells

illumina

2

Body Map 2.0 Project - Individual Human Tissues

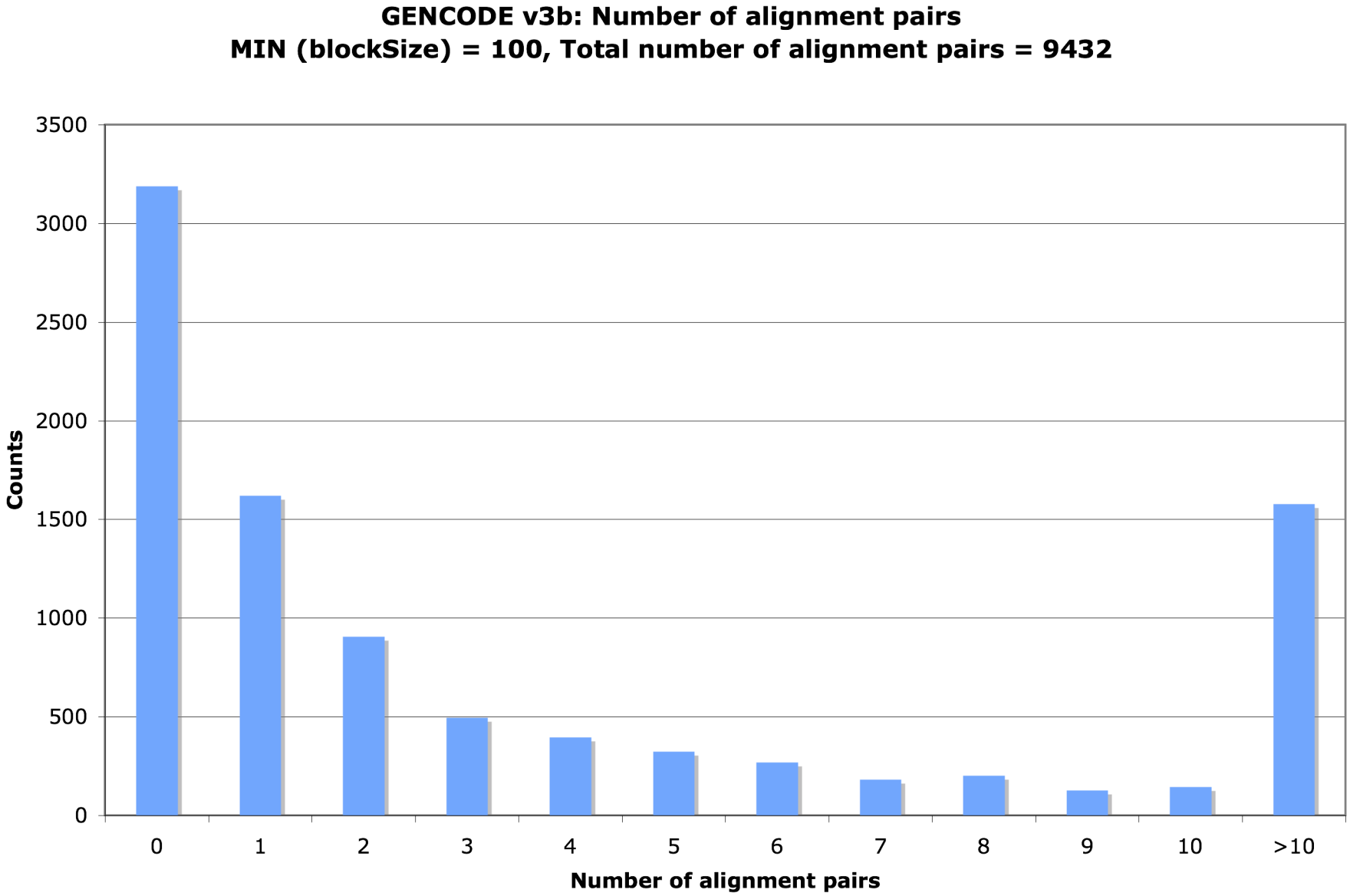
- 16 Different Human Tissues
- Standard mRNA-Seq Library Preps made from poly-A selected mRNA
- Use two different read lengths
  - One run of 2 X 50 bp Paired-End (PE) data
  - One run of 1 X 75 bp single-read data
- Collected one lane of HiSeq 2000 Data per tissue, per run

illumina

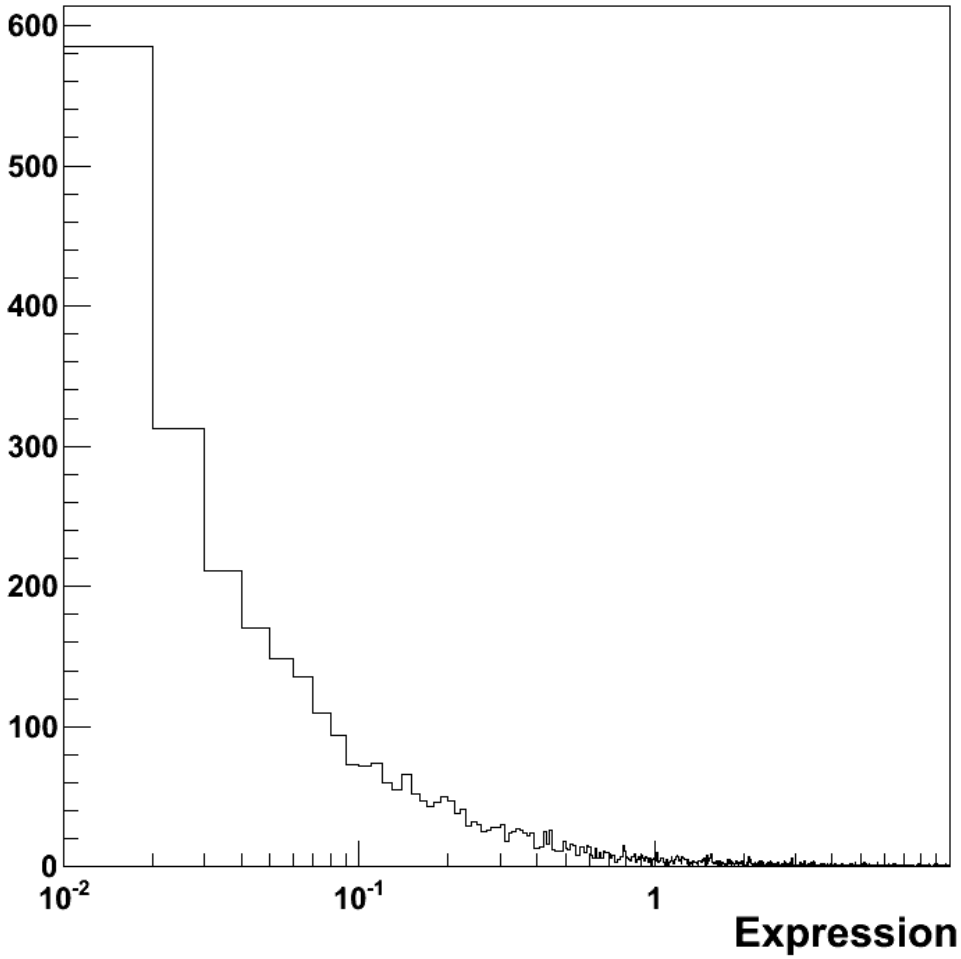
3

Gary Schroth (Illumina, Inc.)

~250 million mapped reads per tissue


GENCODE v3b: Number of alignment pairs
MIN (blockSize) = 100, Total number of alignment pairs = 9432
Counts
3500
3000
2500
2000
1500
1000
500
0
0
1
2
3
4
5
6
7
8
9
10
>10
Number of alignment pairs

Pseudogene transcription: how to set the threshold?



Non-zero pseudogene expression values (DCPMs)

Threshold = 0.1

Alignment pairs	Expressed	Not expressed
0	184	3005
1 to 10	1720	2945
>10	760	818

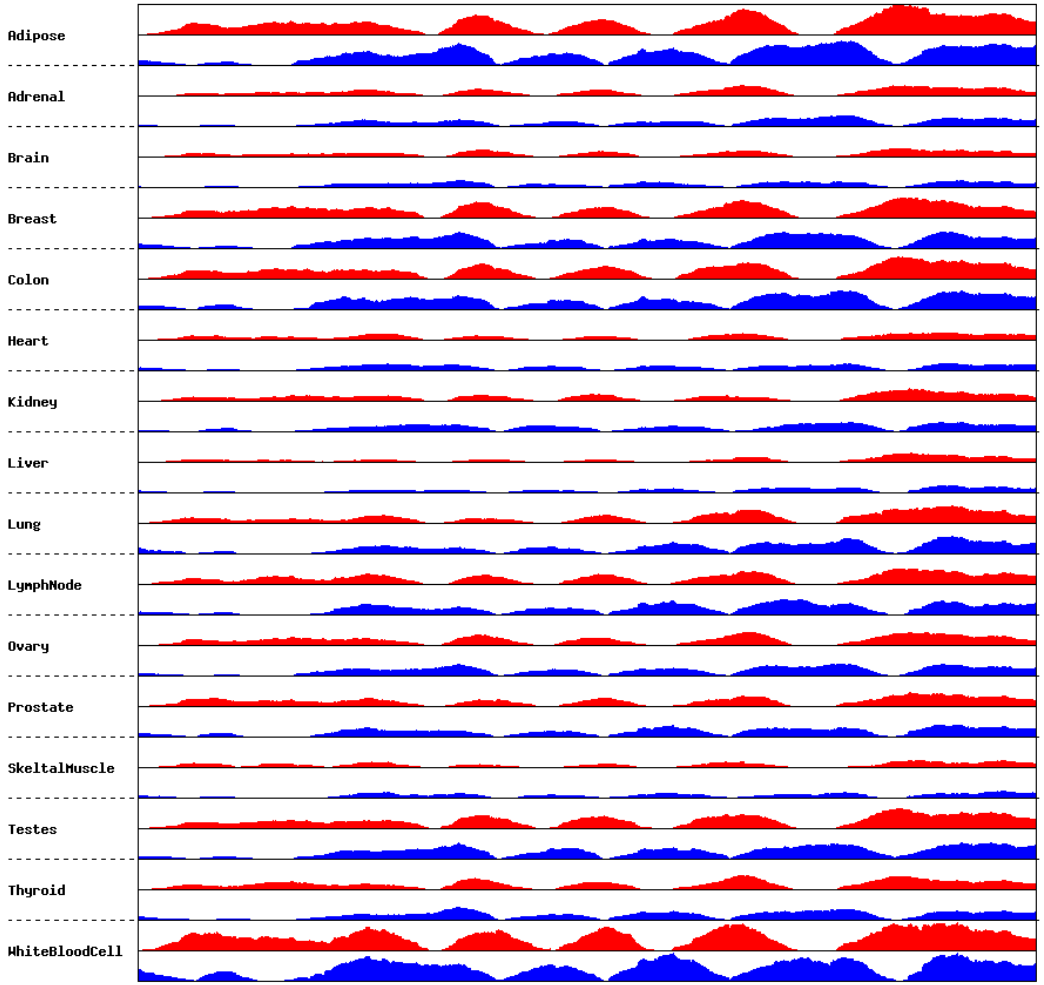
Notes:

- Final number; considered expressed
- Apply PseudoSeq (evaluate all alignment pairs)
  - Expression (pgene) >> Expression (gene)
  - Discordant expression pattern
  - Concordant expression pattern

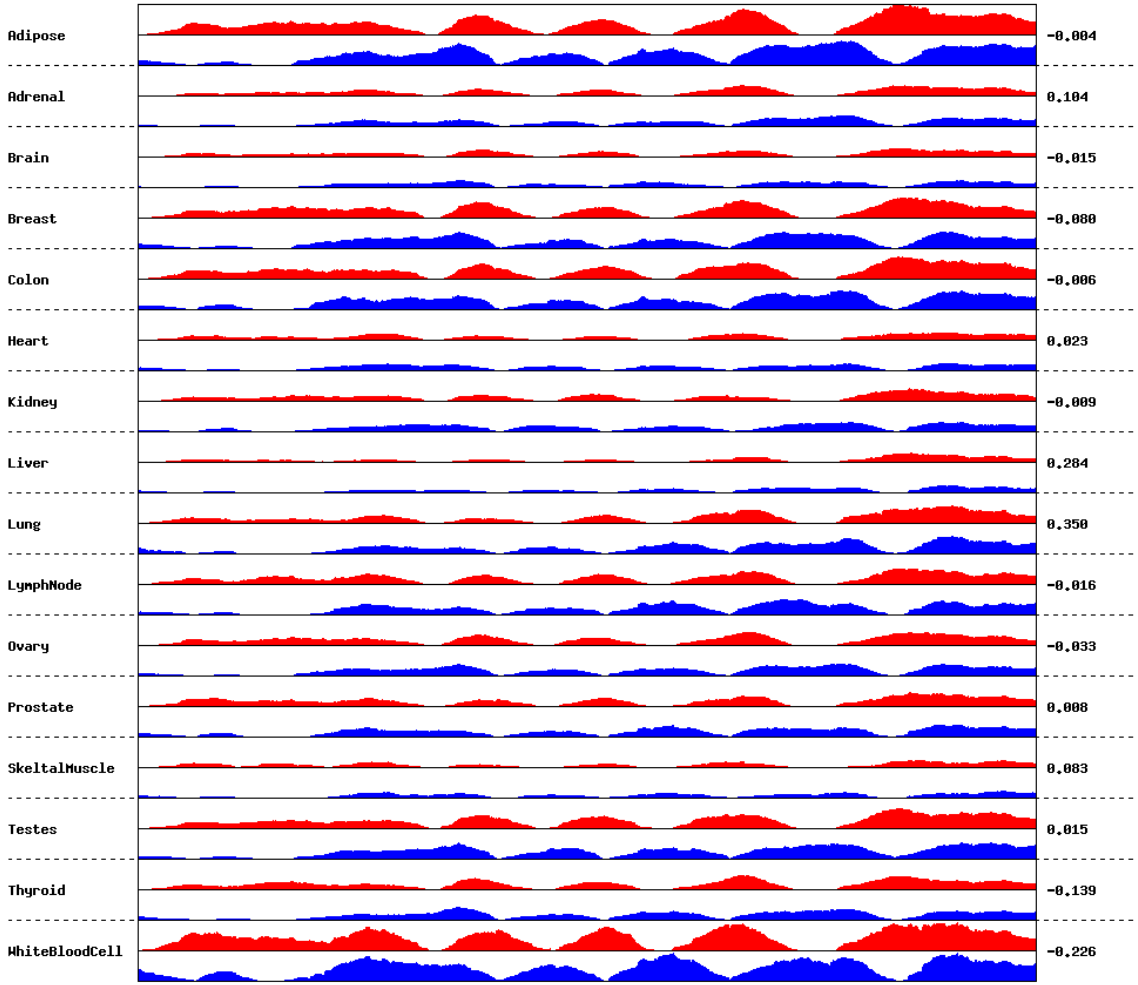
} Considered expressed

- Too many alignment pairs; evaluation not feasible

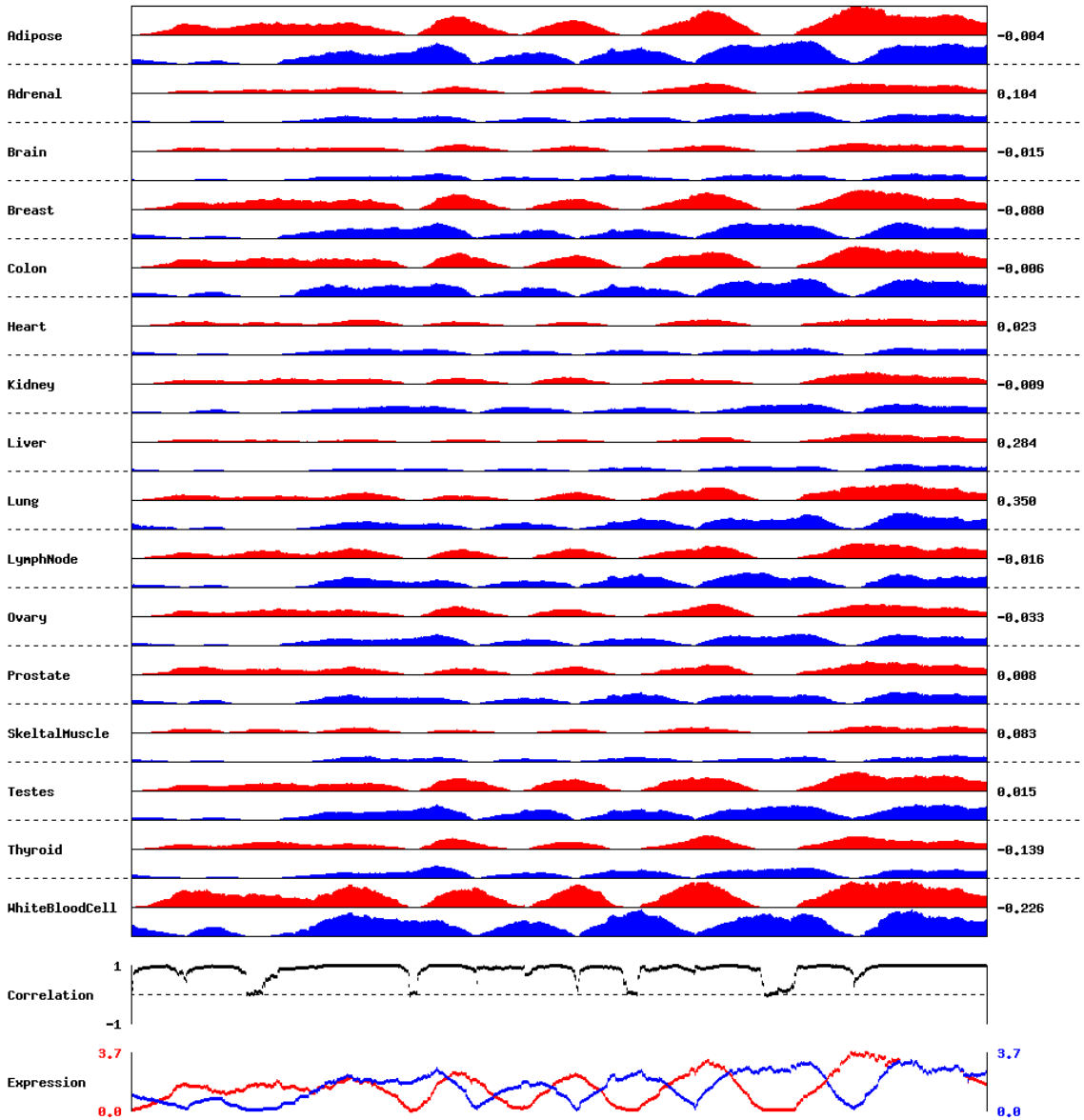
A famous PGENE



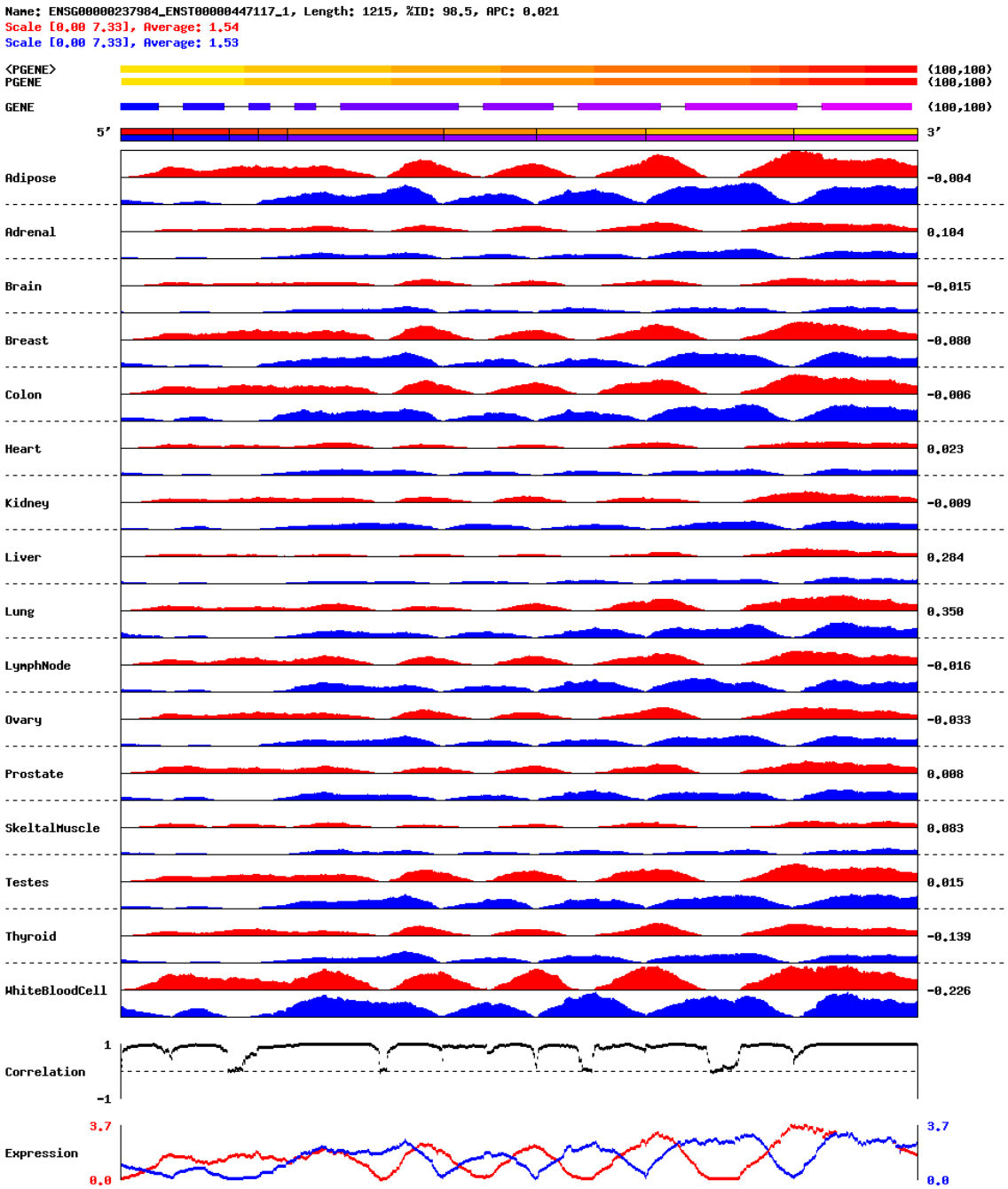
A famous PGENE

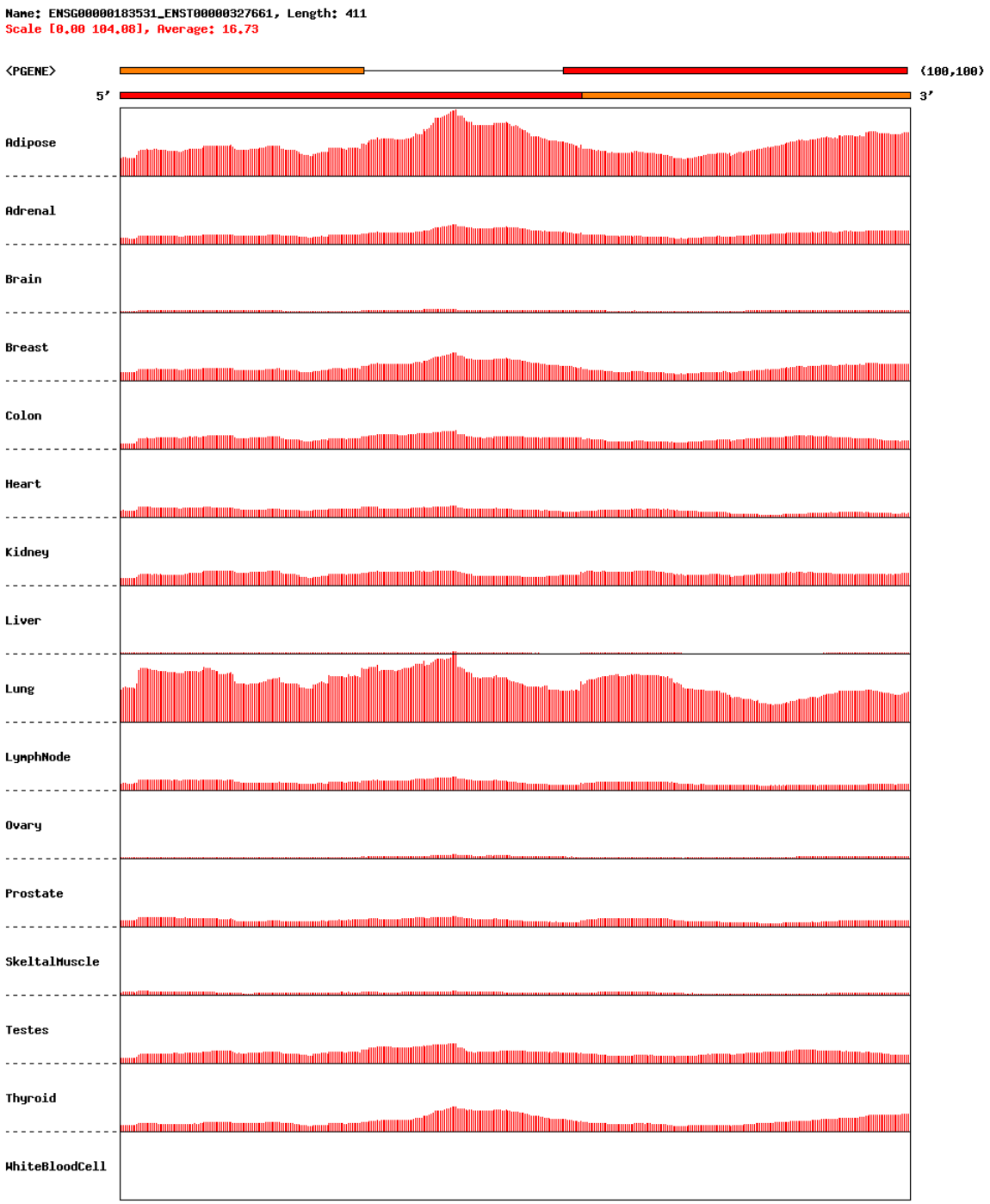


A famous PGENE



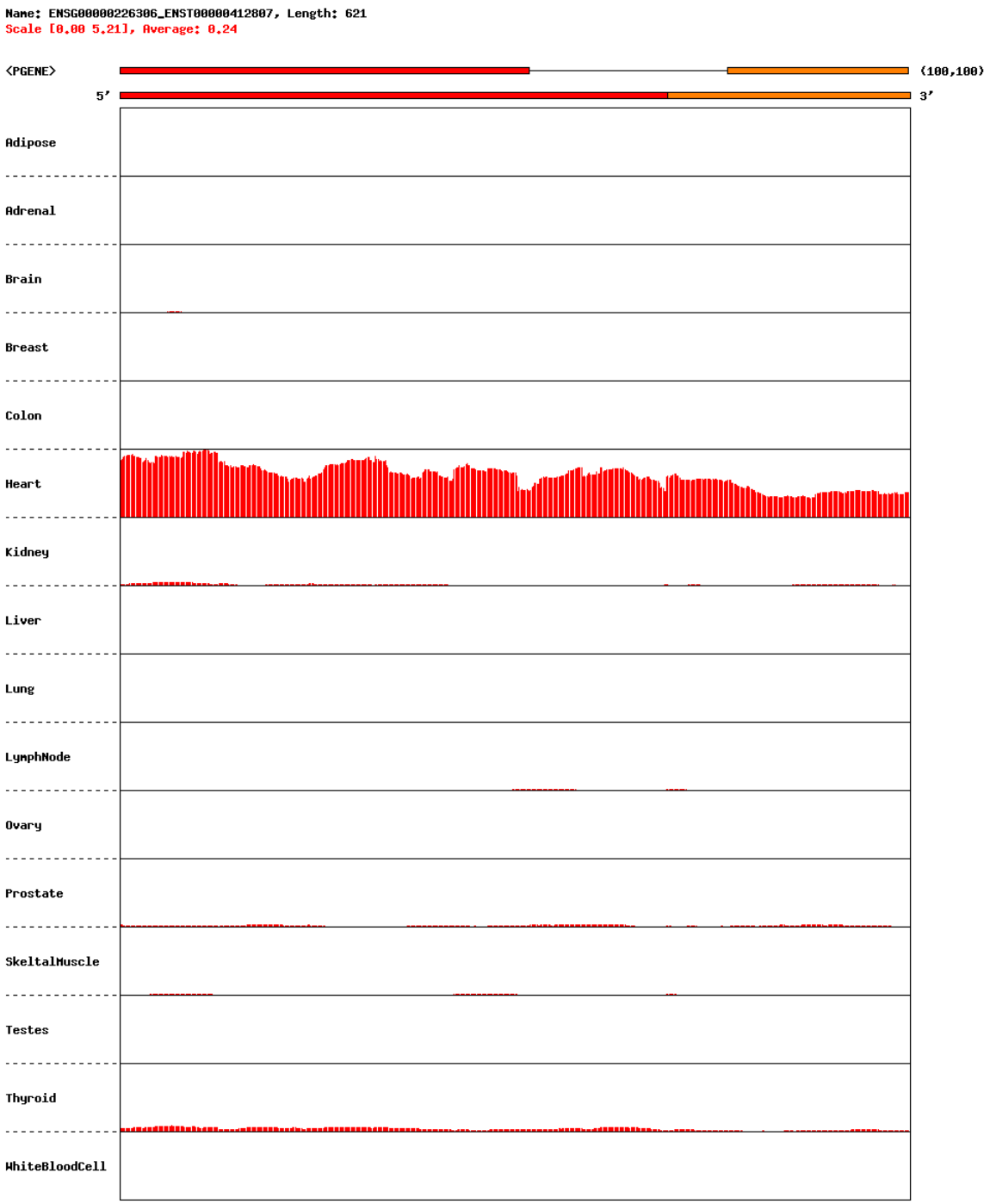
A famous PGENE





Transcribed Pseudogene

Example #1



Transcribed Pseudogene

Example #2

Future steps

- Remapping of the reads
- Analyze the impact of the read mapping procedure
- Calculate global statistics
- Improve scoring
- Improve web-interface

Acknowledgements

VAT
- Suganthi Balasubramanian
- Ekta Khurana

IndelSeq
- Andrea Sboner
- Francesca Demichelis
- Mark Rubin

PseudoSeq
- Andrea Sboner
- Joel Rozowsky

- Mark Gerstein
- Michael Snyder

- genome-tech
- genome-annotation