# Defining the transcriptome of nine vertebrate genomes using RNAseq.

Kerstin Lindblad-Toh, Federica diPalma and Eric Lander

Broad Institute of MIT and Harvard, Cambridge, MA

**To the biomedical research community at large, understanding the content of the genome, particularly the genes, is critically important. The coding and non-coding transcriptomes of both human and mouse have been exquisitely refined through the use of massive amounts of species-specific cDNA sequence data, manual curation and recently mRNA sequencing. However, all other vertebrate genome annotations are largely based on homology to human and mouse genes, due to the lack of cDNA sequence available from other vertebrate species. With the advent of the lower-cost RNAseq technology, the time is right to invest in RNA sequencing from other vertebrate species to improve genome annotations and understand both the coding and non-coding transcriptomes. The Broad Institute currently has laboratory and computational methods in use and under further development to generate high quality RNAseq data for genome annotation. We propose to use <0.02% of NHGRI's annual sequencing capacity at the Broad to improve nine important genome annotations (dog, opossum, bushbaby, rabbit, guinea pig, elephant, armadillo, anolis, and stickleback). At this time, a small investment in RNA sequencing would provide a huge improvement in vertebrate genome annotation, enabling better research across a multitude of biomedical disciplines.**

## Background

Genome annotation is a crucial part of every NHGRI funded genome project – and therefore in the past NHGRI has sometimes funded EST and cDNA sequencing to enable Ensembl to annotate all sequenced vertebrate genomes[1]. It is hard to overemphasize the importance to the scientific community at large of understanding the complete transcriptome and other features in a genome. For genome scientists, gene identification and comparative analysis forms the basis of large number of evolutionary discoveries. To the greater scientific community of organismal biologists, who are the most numerous end-users of genome projects, genome annotation is by-far the most important aspect of a genome project, and for them, the utility of a genome project is directly proportional to the accuracy and comprehensiveness of its gene annotation. However, even though the limited cDNA sequencing already funded and completed has resulted in basic annotations of all recently sequenced genomes, these annotations have been highly dependent on homology to other organisms, and of course, pale in comparison with the detail and specificity of the human and mouse annotations[2][3][4][5]. For some closely related species annotation works well using cDNA sequences from related species but, in many cases within the vertebrate, and also within the mammalian tree, the distances are large enough that the annotation becomes lacking, for example through the inability to place short exons on the genome.

We believe the newly developed RNAseq technology provides the perfect opportunity to improve existing genome annotations at a reasonable price, and thereby will greatly increase the utility of those genomes to the scientific community. RNA-seq by Illumina is orders of magnitudes cheaper than the previous standard – cDNA library construction followed by Sanger sequencing. This makes annotation improvement of a cohort of existing genomes feasible today. The critical technologies required to make RNAseq practical for use in annotation have already been developed at the Broad Institute and are being further improved through both laboratory methods and analytical algorithms. These technologies include normalization of Illumina RNAseq libraries to ensure the capture of low-expression transcripts, strand-specificity of those libraries to determine the strandedness of transcripts sequenced, and the informatic ability to assemble Illumina RNAseq data into full-length transcripts and to align those full-length transcripts, as well as the Illumina reads themselves, to genome assemblies.

With the improved sequencing efficiency of RNAseq, has come the ability to identify not only protein coding genes, but also to detect multiple uncharacterized, abundant non-coding RNAs including lincRNAs and various types of small RNA. For many species virtually nothing is known about these kinds of RNAs. We therefore propose performing RNAseq from nine selected species previously sequenced by the Broad with the aim of generating a near complete automated annotation of protein coding genes as well as to provide information for a large number of non-coding RNAs.

[1] Curwen, V. et al. "The Ensembl automatic gene annotation system," **Genome Research**, 2004, 14(5): 942-50.

[2] International Chicken Genome Sequencing Consortium. "Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution," **Nature**, 2004, 432 (7018):695-716.

[3] Mikkelsen, T.S. et al. "Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences," **Nature**, 2007, 447 (7141):167-77.

[4] Warren, W.C. et al. "Genome analysis of the platypus reveals unique signatures of evolution," **Nature**, 2008, 453(7192): 175-83.

[5] Clamp, M. et al. "Distinguishing protein-coding and noncoding genes in the human genome," **PNAS**, 2007, 104(49): 19428-33.

## Species choice rationale

Annotation is necessary and annotation improvement is tremendously useful for *any* sequenced genome, but there are also specific reasons why these nine species could especially benefit from improved annotations (see Table 1). Primarily the species were selected either for their role as model organisms or as the first sequenced genome of a novel clade. The current annotation of these genomes contain between 11,000-15,000 gene predictions, while most of these genomes should have roughly 20,000+ genes. For some species, such as the dog, additional efforts have identified a somewhat improved gene set of ~19,700 genes using comparative genomics and other approaches, but gene models are still incomplete, for example missing first exons and with abundant errors. Thus we propose generating RNAseq data from multiple tissues to produce a much improved, albeit not manually curated annotation.

**Table 1: Species proposed for RNAseq and their current ENSEMBL gene content compared to that of human.**

| Species | Protein-coding genes in current annotation | Protein-coding genes with species specific support | Specific reason to improve |
|---|---|---|---|
| **Human** | ~21,000 | ~21,000 | |
| **Dog** | 13,527 | 2,504 | Biomedical model |
| **Rabbit** | 11,622 | 495 | Immunological model |
| **Guinea Pig** | 14,232 | 531 | Immunological model |
| **Elephant** | 13,168 | 35 | Afrotheria branch representative |
| **Armadillo** | 12,123 | 0 | Xenarthra branch representative |
| **Bushbaby** | 13,214 | 12 | Distant primate and model |
| **Opossum** | 12,863 | 542 | First marsupial sequenced |
| **Anole Lizard** | 12,084 | 36 | First lizard sequenced |
| **Stickleback** | 15,078 | 71 | Genetics model organism |

Some more specific justification:

- The dog is an important model for a wide variety of biomedical research, such as cancer, diabetes and epilepsy. It is particularly useful for whole genome association studies due to its unique population structure, resulting from its domestication bottleneck, and bottlenecks from recent breed creation.
- Rabbit and guinea pig are significant models in immunological research. However, the vast majority of the gene models in their genome annotations are based on homology to mouse and human genes, which impedes immunological research in particular due to its sensitivity to protein sequence precision.
- Opossum, bushbaby, elephant and armadillo have smaller biomedical research communities, but their genomes are especially valuable to the study of comparative genomics, due to their crucial positions on the mammalian phylogenetic tree. As with the rabbit and guinea pig, these organisms base the overwhelming majority of their gene models on homology with mouse and human genes, considerably confounding efforts to understand gene evolution.
- The green anole lizard and the stickleback fish also have biomedical and evolutionary biology scientific communities that depend on a proper annotation for their research. But lizard and stickleback scientists suffer even greater difficulties than those who study mammals, since the lizard and stickleback genomes are considerably further from their nearest sequenced genome than any of the mammals listed here, thus making their gene annotations even less accurate. The stickleback is a particularly good candidate for genome annotation improvement, since the quality of the stickleback assembly is much higher than that of other sequenced fish.

Although this is just a selection of genomes, many other vertebrate genome annotations will have some benefit from improving these nine genome annotations, as there will be more data at a closer evolutionary distance to use for homology-based annotations than what is currently available.

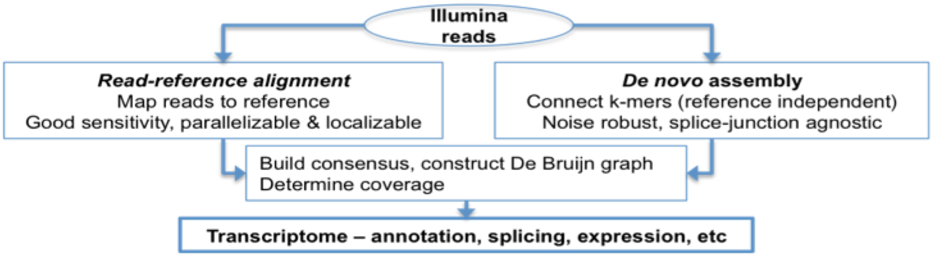
## Sequencing Strategy

We propose to sequence RNA from 11 different tissues (brain, heart, skeletal muscle, testis, ovary, whole embryo, blood, skin, lungs, liver and kidney) in 9 different species (dog, opossum, bushbaby, rabbit, guinea pig, elephant, armadillo, anolis, stickleback). It will be more challenging to obtain tissues from species such as bushbaby and elephant, and so we will sequence more opportunistically from those species – obtaining and sequencing tissues whenever possible. It is also unlikely that we will be able to obtain embryonic tissue from those species. We will utilize the best available Illumina protocol for mRNA sequencing of polyA RNA and generate a library for each tissue in each species (Yassour et al., 2009, Levin et al., 2009). When possible we will also make a microRNA library and sequence (Pfeffer et al., 2005).

We propose to sequence each tissue-specific sample separately, instead of pooling them, to increase the likelihood of capturing low-copy transcripts and splice-forms through the use of our normalization techniques. The depth of coverage of each species' transcriptome will depend heavily on the degree of normalization. The Broad Institute normalizes RNA-seq Illumina libraries using the duplex-specific nuclease (dsn). Normalization protocols are currently under development to increase efficiency and reduce input RNA quantity requirements. We also have strand-specific RNA-seq protocols in process (Parkhomchuk et al., 2009, Yassour et al., 2009, Levin et al., 2009), which will be very useful for genome annotation improvement, as they define which strand of a DNA duplex is transcribed. Our eventual goal is to be able to combine normalization and strand-specificity into a single library protocol. In the basic protocol a normalized libraries will be generated, but when possible an strand-specific unnormalized library will also be made to allow more careful quantification of transcripts as well as to allow the detection anti-sense transcripts.

Analysis

The Broad Institute is currently developing two main RNA-seq analysis methods: alignment of reads to the genome followed by transcript assembly (SCRIPTURE (Guttman et al., 2010), and assembly of reads independently of a genome reference *ANANAS* (manuscript in preparation).


Illumina reads
*Read-reference alignment*
Map reads to reference
Good sensitivity, parallelizable & localizable
*De novo* assembly
Connect k-mers (reference independent)
Noise robust, splice-junction agnostic
Build consensus, construct De Bruijn graph
Determine coverage
Transcriptome – annotation, splicing, expression, etc

Our alignment-first method, SCRIPTURE, is a statistical transcriptome reconstruction method that takes advantage of longer Illumina reads to build a graph whose edges come from all reads that align to the genome, spanning exon-exon junctions. Our assembly-first method, ANANAS, builds consensus cDNAs in a noise-robust fashion, independently of a reference genome sequence. It is being developed to identify alternative splice forms, as well as serve as an input to annotation pipelines where the reference sequence is incomplete. For example, assembling the reads before genome alignment would be especially useful in the case of dog RNA-seq data, since we expect the assembled cDNAs to contain first exons missing from the actual dog genome assembly.   These methods will be further developed and combined into a mixed method that optimally uses the strengths of both methods.

All data generated (reads and transcripts) will be made available to Ensembl and other genome browsers to ensure the data can be utilized in vertebrate annotation. Ensembl is prepared to take both assembled cDNA sequences and individual reads as inputs to their annotation pipeline.  We will be working with Ensembl to ensure they can utilize our data in the best possible way for each different species/datasets.

The Broad Institute is also currently developing tools to improve all the algorithms as well as the tools that will utilize these data to annotate and serve the data on the IGV genome browser (separate proposal to be submitted October 5th, 2010).

**Time line**

We expect to perform the majority of the sample collection (except for samples that are only opportunistically available), sequencing and transcript assembly in the coming year, but expect that some of the annotation efforts will take longer to complete.

## Budget

We expect to generate ~1,000 Gb of data, totaling <0.02% of the annual NHGRI capacity at the Broad Institute.

## Reference

**Mitchell Guttman**, Manuel Garber, Joshua Z Levin, Julie Donaghey, James Robinson, Xian Adiconis, Lin Fan, Magdalena J Koziol, Andreas Gnirke, Chad Nusbaum, John L Rinn, Eric S Lander & Aviv Regev. *Ab initio* reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology* 28, 503-510.

**Levin JZ**, Berger MF, Adiconis X, Rogov P, Melnikov A, Fennell T, Nusbaum C, Garraway LA, Gnirke A. (2009) Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. *Genome Biol*. 10(10):R115.

**Parkhomchuk, D.,** Borodina, T., Amstislavskiy, V., Banaru, M., Hallen, L., Krobitsch, S., Lehrach, H. and Soldatov, A. (2009) Transcriptome analysis by strand-specific sequencing
of complementary DNA. *Nucleic Acids Res*. 37,e123.

**Pfeffer S**, Lagos-Quintana and Tuschl, Cloning of small RNA molecules. Curr Protoc Mol Biol. 2005 Nov;Chapter 26:Unit 26.4.

**Yassour, M.,** Kaplan, T., Fraser, H. B., Levin, J. Z., Pfiffner, J., Adiconis, X., Schroth, G., Luo,S., Khrebtukova, I., Gnirke, A., Nusbaum, C., Thompson, D. A., Friedman, N. and Regev, A. (2009) Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proc Natl Acad Sci U S A*. 106,3264-9.