UNIVERSITY OF CALIFORNIA
SANTA CRUZ

COMPARATIVE ANNOTATION TOOLKIT (CAT) - SIMULTANEOUS CLADE AND PERSONAL GENOME ANNOTATION

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOINFORMATICS

by

Ian Fiddes

December 2017

The Dissertation of Ian Fiddes
is approved:

Professor David Haussler, Chair

Professor Benedict Paten

Thomas Keane, Ph.D.

Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Comparative Annotation Toolkit (CAT) - simultaneous clade and personal genome annotation

by

Ian Fiddes

The recent introductions of low-cost, long-read and read-cloud sequencing technologies coupled with intense efforts to develop efficient algorithms have made affordable, high-quality *de-novo* sequence assembly a realistic proposition. The result is an explosion of new, ultra contiguous genome assemblies. To compare these genomes we need robust methods for genome annotation. I describe the fully open source Comparative Annotation Toolkit (CAT), which provides a flexible way to simultaneously annotate entire clades and identify orthology relationships. I show that CAT can be used to improve annotations on the rat genome, annotate the great apes, annotate a diverse set of mammals, and annotate personal, diploid human genomes. I demonstrate the resulting discovery of novel genes, isoforms and structural variants, even in genomes as well studied as the rat and great apes, and how these annotations improve cross-species RNA expression experiments.

Acknowledgments

In addition to my committee members, I would like to thank the following people for their advice and support:

Family Members

- Sheela Fiddes
- John Fiddes
- Karen Talmadge

Lab Members

- Sofie Salama
- Joel Armstrong
- Mark Diekhans
- Kristof Tigyi

Collaborators

- Evan Eichler
- Zev Kronenberg
- David Gordon
- Jason Underwood
- Anthony Doran
- Jingtao Lilue
- Cristina Sisu
- Fergal Martin
- Paul Flicek
- Ron Korstanje
- Anuj Srivastava
- Laura Reinholdt
- Valerie Schneider
- Vishal Kumar Sarsani
- Lukas Kuderna
- Tomas Marques

PhD Cohort

- Arjun Rao
- John Vivian
- Ed Rice
- Nathan Schaefer

Chapter 1

Genome Alignment And Comparative Annotation

This section provides an overview of the history of whole genome alignment and comparative annotation, and leads into why the development of CAT is essential to the new era of comparative genomics. This comprises a review paper on the subject being written by myself and Joel Armstrong.

1.1 Whole-genome alignments

1.1.1 Introduction

Alignment is possibly the most fundamental problem in genomics. The alignment problem is to establish a mapping between the letters of a set of sequences that approximates some relation that the user is interested in. In comparative genomics, we are generally interested in the *homology* relation—that is, does the lineage of two bases coalesce at a single base in a single organism at some point in time? In typical real-world comparative genomics, there is no clear proof of homology, as we have absolutely no access to the true history of every base in a set of sequences. However, we can use our knowledge of molecular evolution to construct very good approximations to the homology relation. The potential for using sequence similarity to approximate homology was recognized and applied very early on, starting with the pioneering work of Needleman and Wunsch on optimal pairwise global alignment [1]. The pairwise global alignment

work was quickly specialized to perform *local alignment*, which calculates the optimal alignment of subsequences rather than sequences, by Smith and Waterman [2].

The *whole-genome alignment* problem involves aligning two or more genomes, or other large sequences or set of sequences, together. While whole-genome alignment is essentially just alignment writ large, it is generally treated differently than short sequence alignment because two factors not usually considered in short alignments become impossible to ignore: size and rearrangements. The size of the problem in whole-genome alignment causes alignments to take too long to be practical, forcing efficiency considerations to be taken into account. The traditional dynamic-programming algorithms require $O(nm)$ time and space, where n and m are the lengths of the two sequences; obviously, as n and m grow to genome-scale the problem becomes too expensive to solve in practice. Another consideration is how genome rearrangements complicate the alignment problem. Smith-Waterman and Needleman-Wunsch both produce alignments which have *fixed order-and-orientation*, that is, insertions, deletions, and substitutions are the only allowed edit operations. When looking at short or well-conserved sequences, like genes, this requirement is usually fulfilled. But at large evolutionary distances, genomes almost always contain more complex rearrangements with respect to each other—inversions, transpositions, and duplications all cause breaks in order and orientation that cannot be captured under constant order and orientation.

One obvious solution to these problems is to use a fast approximate local alignment algorithm (like BLAST [3]) and simply use the collection of all local alignments that it finds as the whole-genome alignment. However, the nave local alignment approach has its own problems. Whole-genome local alignments have both too low sensitivity and too low specificity to be useful at substantial evolutionary distances. That is, local alignments will miss homologous sequence that, by chance, happened to be further diverged. They will also capture spurious alignments that can obscure the more useful data. Even when they correctly identify homologous regions, the end-user is more often interested in *orthology* rather than homology: ancient duplications may share similar sequence, but often do not share similar function. We call any alignment that allows rearrangements (i.e. does not have fixed order-and-orientation) and attempts to determine orthology rather than just homology (even if restricted to single-copy) a

whole-genome alignment (or for short, a *genome alignment*). Most whole-genome alignment methods are based on local alignments, but do some filtering and post-processing to construct a useful end product [4].

As long DNA sequences became available, it was soon recognized that Needleman-Wunsch or Smith-Waterman alignments were far too slow to be useful for megabase-scale sequences, much less chromosome-scale sequences. The impractical running time of global alignment drove the development of several tools [5, 6, 7] that produce an approximately optimal global alignment through the use of high-confidence *anchors* in a single order and orientation, which are then used to partition the alignment into smaller problems which can be more efficiently solved. These anchors provided a very efficient and reliable way to break up the alignment problem, but relied on a constant order and orientation, which excluded any possibility of noticing rearrangements.

Chaining [8] is a powerful technique for making sense of pairwise local alignments. Chains are simply maximal combinations of local alignments that maintain a single order and orientation. Chaining provides a good way of filtering out spurious alignments, which are likely to form short, low-scoring chains. However, the set of chains can often include distant paralogs or spurious sequence, which makes it difficult to understand the rearrangements that have taken place between the two input genomes. *Netting* [8] is a related technique that makes rearrangements relative to a reference genome much easier to find. In essence, netting finds the best-scoring set of chains that covers the bases of the reference genome only once. This makes it very easy to find high-confidence rearrangements like transpositions, inversions, and deletions, but throws away the effect of duplications in the target genome.

Choosing the single "best" target alignment for each region, which we will call the *single-copy* strategy, is a common way [9, 10] to deal with the problems that duplications cause. However, the best-fit strategy will not always find a correct ortholog, and indeed even reciprocally-best-fit is not enough to guarantee finding an ortholog [11]. But most importantly, lineage-specific duplications should not be ignored. When lineage-specific duplication occurs, a gene outside that lineage will have *multiple* orthologs in the lineage, and should be aligned to multiple copies [12]. Single-copy alignments implicitly assume that orthology is a one-to-one relationship. However, in nature, orthology is

a many-to-many relationship [12]. When that assumption of one-to-one orthology is violated, single-copy alignments can be very misleading.

1.1.2 Multiple alignment

Often it is necessary to consider the alignment between a set of more than two sequences, which we call *multiple alignment*. A multiple alignment is defined as an equivalence relation $\sim$ on a set of sequences $\mathcal{S} = \{s_1, s_2, \ldots\}$, such that for two bases $b_1 \in s_1 \in \mathcal{S}$ and $b_2 \in s_2 \in \mathcal{S}$, $b_1 \sim b_2$ if they are considered to be aligned to each other. The alignment is partitioned into *columns* by the equivalence classes of $\sim$: i.e. every base is related to all bases in its column, and no two bases in different columns are related. Unfortunately, multiple alignment is a significantly more difficult problem than pairwise alignment. Finding an optimal multiple alignment, even using very simple objective functions, has long been known to be NP-hard [13]. Heuristics must be employed to efficiently solve the multiple alignment problem. *Progressive alignment* is the most popular strategy for approximate multiple alignment [14]. Progressive alignment uses as an additional input a *guide tree* relating the input sequences. The most closely related sequences are aligned first, then the resulting alignment is itself aligned to other sequences or alignments, following the structure of the guide tree. Often consensus sequences are used as a method of aligning alignments. Since the multiple alignment problem is so difficult, another approach is to use a single *reference* sequence to base the alignment on. All other sequences in the multiple alignment are simply aligned to this genome in a pairwise fashion, then the several pairwise alignments are combined to form a *reference-biased* multiple alignment. This approach performs very well when viewed from the reference genome, but information relating genomes distant from the reference is lost. See 1.1 for an illustration of this effect. In the mid- to late-2000s the first methods for reference-free multiple genome alignment allowing multiple copies began to appear (notably the Enredo-Pecan-Ortheus (EPO) pipeline [15] and the A-Bruijn aligner [16]). The EPO pipeline especially began to see wide use as part of the Ensembl genome browser [17]. While impressive, these pipelines left significant room for improvement, especially with regard to finding small-scale order-and-orientation-breaking rearrangements [15].

1.1.3 Reference-free alignment

1.1.3.1 Genome histories

Alignments are conventionally described as a set of columns, each containing a set of bases that are all related to each other by some alignment relation $\sim$. Usually this relation represents orthology rather than homology. However, in that case, this model falls apart when considering reference-free alignments with multiple copies per genome. The orthology relation is not transitively closed [12], so it is impossible in the general case to create a set of columns containing bases that are all orthologous to each other. The only way to represent a reference-free, multi-copy, orthologous multiple genome alignment is by associating the alignment with phylogenetic trees, which are inferred (even if implicitly) during the alignment process. We term these types of alignments *genome histories* to reflect that they require a different representation than typical alignments (which can be represented by a collection of only blocks or columns).

A *genome history* $\{\mathcal{S}, \sim, T_c, t_s, L\}$ consists of a set of genomes $\mathcal{S}$, a multiple alignment $\sim$ relating the bases of those genomes, a reconciled tree $t \in T_c$ for each column in that alignment, a species tree t_s, and, optionally, a set of *links* L between columns, indicating the ordering of the ancestral chromosomes. The columns of the genome history reflect the *homology* rather than *orthology* relation. Since homology is transitive, the homology-based alignment can be represented by columns. The set of trees (hereafter referred to as *column trees*) indicate the evolutionary history of the bases in each column. Where there are duplications, gains, or losses, the column tree $t \in T_c$ will differ from the species tree t_s. Though the genome history representation we present here is not the only possible representation, any other representation (such as a collection of all pairwise orthology relationships) can be transformed into this one.

A genome history can be used to define both *orthology* and *paralogy* relations. The orthology relation, which we will symbolize by $\sim_o$, uses the column trees of the genome history to determine which of the homologous bases in a column are also orthologous to each other. The orthologous bases are those homologous bases whose lineage coalesces in a speciation event in the column tree [12]. The paralogy relation $\sim_p$ simply relates homologous bases which are not orthologs.

A genome history can be *projected* onto any genome to create a more conven-

tional referenced multiple alignment. These projected, reference-based alignments are collections of columns, each containing exactly one reference base, where every base in the column is orthologous to the reference base, but *not* necessarily orthologous to every other base in the column. These projected alignments are useful because they can be represented in conventional formats like MAF, and used as input to existing analysis tools.

1.1.3.2 HAL format

One difficulty with producing reference-free alignments is that conventional text-based alignment formats like MAF cannot be efficiently randomly accessed from every genome. The Hierarchical Alignment Format [18] (HAL) was designed to be an efficiently accessible format representing a genome history, including any ancestral reconstructions available.

HAL allows projection from this genome history onto any reference genome (including ancestors), creating a multiple genome alignment showing what is orthologous (related by $\sim_o$) to every base in that genome. This projection can be output in a traditional format like MAF, or simply used on-demand to visualize the alignment [19] or as part of downstream analysis pipelines.

1.1.4 Local alignment tools

Because genome alignment tools rely heavily on local alignments of some form, local alignment tools play a large role in genome alignments. Since finding optimal local alignments has prohibitive time and memory requirements, approximate local aligners in the vein of BLAST [3] are used almost exclusively. BLASTZ [9] and its successor LASTZ [20] are local aligners tuned to be more sensitive than normal BLAST, using PatternHunter-esque spaced seeds [21] while also allowing transitions. LAST [22]

1.1.5 Multiple genome alignment pipelines

An overview of popular or historically significant alignment multiple genome alignment tools is given in **Table 1.5**. In the following sections, we detail some of the most significant pipelines that are still in wide use.

1.1.5.1 MultiZ

MultiZ [10] is a reference-biased multiple genome alignment tool originally developed as part of the TBA [10] program. Because TBA is restricted to producing colinear multiple alignments (though there exists an unpublished version that removes that restriction), MultiZ sees much wider use than TBA itself. It is the tool currently used to generate the multiple alignments on the UCSC Genome Browser [23].

MultiZ, technically speaking, is just a method of aligning alignments. When MultiZ alignments are produced, usually pairwise alignments from a given reference to all other species are generated using a local alignment tool, sometimes post-processed using chains and nets, and then the "autoMZ" command is used to progressively align together these pairwise alignments using a guide tree.

1.1.5.2 EPO

The Enredo-Pecan-Ortheus (EPO) pipeline [15, 24] is a reference-free multiple alignment pipeline that, unlike TBA, can handle rearrangements. It is in wide use, being one of the main multiple alignments available on the Ensembl genome browser [25]. The process begins with a relatively sparse set of anchor points that are known homologies within a set of genomes. The Enredo algorithm builds a sequence graph from these anchors, and through various operations, attempts to remove homologies that are likely to be spurious or uninteresting. The Pecan algorithm then fills in the gaps between the sparse anchors selected by the Enredo algorithm. The Ortheus algorithm [24] is then optionally run to generate ancestral sequences for all blocks, creating a genome history. While EPO is in principle reference-free, the method that is currently used to generate its anchors is reference-biased [15].

1.1.5.3 Cactus

Cactus uses an overall strategy similar in principle to the anchoring approach described above. The notion of a *cactus graph* [26] is used to create a filtered, high-confidence set of anchors. The unaligned space between anchors is then aligned using a sensitive pair-HMM to create a final multiple alignment. The first step of the Cactus process is to take small, uncertain local alignments captured by LASTZ [20] (which is

similar to BLAST [3]), and combine them navely to create a multiple alignment. Given the typical evolutionary distances involved, LASTZ is tuned to be very sensitive, but not very precise. The low precision means that the local alignments may be spurious (a small seed happened to match, and happened to be extended, in a region which is not truly homologous). The local alignments may also conflict—that is, several alignments may disagree on how to align a particular region. These inconsistencies and spurious alignments will manifest as tiny rearrangements—breaks in order and orientation—in the alignment. Using the Cactus Alignment Filter (CAF) algorithm defined in [27], these small rearrangements, which are unlikely to be biological, in the multiple alignment are discovered and removed, producing an alignment that only contains rearrangements longer than a certain length. After this process, the cactus graph contains anchors that are very likely to represent true regions of homology, but will have unaligned regions of homology between the anchors, which local alignment was not sensitive enough to pick up, or which were deleted in the CAF process. The Base Alignment Refinement (BAR) process [27] fills in these unaligned but homologous regions.

1.1.5.4 progressiveCactus

The version of Cactus published in 2011 [27] was highly effective at aligning a small number of genomes in the tens to hundreds of megabases [28], but because it scaled quadratically with the total size of all genomes in the alignment problem, it could not efficiently create the alignments we needed, which require us to align hundreds of vertebrate-sized genomes. Recently a progressive-alignment extension (called progressiveCactus) to the original Cactus algorithm has been developed, which can efficiently scale to hundreds of genomes. The progressiveCactus process works as follows. First, the problem is decomposed into several subproblems using an input guide tree. There is one subproblem per internal node in the guide tree. Each subproblem involves aligning several genomes using the traditional Cactus process: the *ingroup* (children of the internal node) and *outgroup* (non-descendants of the internal node) genomes for the subproblem. This subproblem alignment is then used to infer a "reference" assembly that contains all blocks involving an ingroup. The blocks are arranged into sequences according to an algorithm that attempts to maximize the consistency between the order

and orientation of all the sequences in the alignment [29]. The base-level sequence for these blocks is then generated by finding the ML base for each column using the guide tree. This assembly is a reconstruction of the ancestral genome at that node, which functions as a consensus sequence for the ingroups below it. The reference assembly is then fed as input into subproblems further up the guide tree.

1.2 Comparative annotation

1.2.1 Introduction

Genome annotation is the process of finding functional elements in a genome assembly. Generally, these take the form of protein coding genes, but can also include non-coding transcripts[30], chromatin configuration[31], DNase hypersensitivity[32], CpG islands[33], and population variation [34].

The task of automatically annotating genome assemblies has been considered since the first full length genomes were released in the mid-1990s [35, 36, 37]. This task is often divided into two categories — *ab-initio* prediction, or the computational prediction of exon-intron structure using statistical models, and sequence alignment based approaches, which map any of EST, cDNA or protein sequences on to an assembled sequence to discover transcripts [25]. Some annotation pipelines combine both sources of transcript prediction to generate a final annotation set [38, 39].

Recent improvements in sequencing technologies, including long-read [40] and linked-read technology [41], have provided the ability to produce high quality genome assemblies at prices that make genome assembly an affordable experiment to labs across the world. This has led to the formation of consortia that aim to produce genome assemblies on a wide scale, including the Vertebrate Genome Project [42], the 200 Mammals Project (`https://karlssonlab.org/2017/08/03/the-200-mammals-project/`) and Insect 5K [43].

This rapid increase in the availability of high quality genome assemblies necessitates the introduction of automated methods that can scale, and that can leverage the improved phylogenetic information that such an array of assemblies can provide. For example, the 200 Mammals Project is specifically designed to allow for the calculation

of base-level conservation across mammalian evolutionary history. This discriminatory power can be leveraged downstream of the assemblies to improve whole genome alignments as well as annotations, and it provides a framework for annotating true many-to-many orthology relationships instead of the current models that rely heavily on annotating relationships relative to mouse and human.

In this new era of genome assembly, consideration must be given not just to assembly and alignment but also to annotation. Annotation is central to the question of how to utilize this explosion of genome assemblies, and high quality annotation sets with orthology mappings across species will enable a wide range of comparative genomics analyses.

1.2.2 Sequence Based Comparative Annotation

In conjunction with the release of the first mouse draft assembly[44], multiple different tools were created to try and leverage comparative information to human to look for genes, including TWINSCAN [45], SGP [46] and SLAM [47]. **Table 1.1** provides an overview of comparative annotation tools, including those written in the years following the mouse genome assembly. These tools provide probabilistic frameworks which combine established single-genome gene prediction approaches [48, 49] with informant data obtained through genomic alignments to improve gene predictions. Notably, all of these tools work only on pairwise alignments and cannot use information extrinsic to this alignment and the underlying input sequences.

As more vertebrate genomes were sequenced, the need for comparative gene predictors that could use more than one informant species arose. Some of the previous tools were re-engineered, as is the case with N-SCAN [50, 51]. However, N-SCAN predicted only 35% of human genes correctly, and using a multiple sequence alignment was no more accurate than a high quality pairwise alignment [52]. In contrast, CONTRAST [53] remarkably was able to accurately predict 65% of human genes using 11 informant genomes. Prior to CONTRAST, practically all gene prediction tools relied on hidden Markov models (HMM), a generative model, while CONTRAST relied on a discriminative support vector machine (SVM) model. The SVM is used to model coding regions, while an additional model called a conditional random field (CRF) is used to model the

gene structure itself. A CRF can be considered a generalization of a HMM [54].

For the most part, after the initial mouse genome project, none of these tools have been used on full vertebrate genomes. There are two reasons for this. First, these tools require very careful parameter training which must be performed on every genome in the alignment. Second, these tools require evaluating all pairwise comparisons leading to running times quadratic in the number of genomes. Combined with the overall lower efficacy of comparative prediction vs. transcriptome and proteome sequence alignment approaches, this has led the field of comparative gene finding to languish for the past ten years.

1.2.3 Transcriptome Evidence Based Comparative Annotation

None of the annotation programs described above were capable of incorporating extrinsic information, instead relying entirely on sequence composition. In species with sufficient transcriptome data, mapping these data to the assembly generally performs far better at gene finding than the *de-novo* approaches outlined above [55, 56].

In the early 2000s, projects like the Mammalian Gene Collection (MGC) [57] were generating full length cDNA sequences for model organisms. These full length transcripts, in addition to expressed sequence tags (ESTs), were being stored in databases like GenBank [58], supplemented by submissions from labs around the world. Tools were developed to incorporate alignments of these sequences in gene prediction, including N-SCAN_EST [59], GenomeWise [60], and AUGUSTUS [55, 61, 62].

While these tools were developed for annotating single genomes with extrinsic information from the same genome, they can be applied in a comparative fashion. Many species of interest have limited transcriptome data available but are closely related to well annotated species. Examples include mouse vs. rat, and human vs. other great apes. Alignment of related transcript sequences is used in the gene builds produced both by Ensembl [25] and RefSeq [38].

Another approach is to use alignments of protein sequences instead of transcript sequences, which is more robust across long phylogenetic distance. The popular annotation pipeline MAKER2 [39] provides such functionality, and recommends providing protein sequences from at least two related genomes [63].

For more distantly related species, approaches that generate profiles of proteins and protein domains may be used. Databases such as InterPro [64] store precomputed models of protein sequences and motifs that are conserved across long periods of evolutionary history. Genewise [60] can perform gene prediction using a profile-HMM like those stored in InterPro. Additionally, AUGUSTUS-PPX [65] is a extension of the AUGUSTUS annotation program that models protein families and combines them with the existing *ab-initio* model.

1.2.4 Transcript Projection

Transcript projection uses sequence alignments to project the coordinates of an existing annotation in one genome to another genome. This powerful approach leverages high quality annotations in well studied organisms to annotate diverse transcripts in related genomes. Many genes and isoforms are expressed in specific tissue types [66], at specific developmental time points, or only in response to specific environmental conditions [67]. Gathering the data for a new species to fully annotate their transcriptome is thus prohibitively expensive. Transcript projection methods allow this to be bypassed, by making use of the high quality information in well studied organisms[68, 69]. Additionally, traditional *ab-initio* gene finding models rely heavily on the signature of protein coding genes, which limits these models' ability to predict UTR sequences, non-coding RNAs such as lncRNAs, and pseudogenes. Transcript projection methods can be combined with any available extrinsic information either from the genome in question or related genomes and provide the highest quality annotation as a result [55].

The first tool to perform transcript projection was Projector [70], which uses a pair hidden Markov model and models exon-intron structure through a pairwise alignment, similar to how tools like TWINSCAN work. However, Projector can make use of the known gene information in one sequence in the alignment to restrict the probability paths to those that match the known gene. A subsequent tool, Annotation Integrated Resource (AIR) [71], introduced the concept of a splice graph, a directed acyclic graphic structure that represents exons as vertices and introns as edges, where isoforms of a gene are paths through this graph. AIR projects transcripts from a reference genome through a syntenic alignment to score the paths in the graph, reducing

the large number of combinations that are biologically improbable.

1.2.4.1 transMap

transMap [55], first developed in conjunction with improvements to AUGUSTUS to model extrinsic information [62], relies on whole genome alignments to project existing annotations from one genome to the other genome in the alignment. This process is purely arithmetic, but has proven to be immensely helpful at providing extrinsic information to guide AUGUSTUS and improve on purely sequence based prediction. Compared to methods that incorporate EST alignments, transMap provides both full length transcript information as well as isoform information. transMap provided the biggest benefit to specificity in AUGUSTUS predictions in all cases except in cases where the existing cDNA repertoire for the species in question exceeded the quality of the reference [55].

1.2.4.2 CESAR

Coding Exon-Structure Aware Realigner (CESAR) [69] is a tool that projects exons through a whole genome alignment, handling splice site shifts that are a common feature of evolutionary change. CESAR is a straightforward hidden Markov model that takes as input the linear alignment of a single exon to other genomes with a small amount of flanking intronic sequence, and outputs a re-aligned region that accounts for exon frame and evolutionary change. CESAR was able to achieve a nearly 89% accuracy at realigning splice sites, leading to the number of frameshifts seen when mapping human genes to mouse to drop from 2.7% to 0.3%. These spurious frameshifts must be addressed when working with transcript projection methods.

1.2.5 Comparative Augustus

AUGUSTUS recently added a novel objective function parameterization option that makes use of whole genome alignments to predict coding genes simultaneously in every genome in the alignment [72] called Comparative AUGUSTUS or AugustusCGP. With recent updates, training the AugustusCGP model is straight-forward and integrated in the AUGUSTUS binary. In contrast to previous comparative gene finding

tools, AugustusCGP runs linearly in the number of genomes, making the possibility of annotating dozens of genomes computationally tractable. Currently, AugustusCGP relies on a referenced multiple-genome alignment format called MAF, and as such cannot annotate many-to-many relationships.

AugustusCGP makes use of the phylogenetic information inherent in the whole genome alignment to look for evidence of negative selection. This is done by calculating the ratio of synonymous and non-synonymous substitutions in candidate exons. AugustusCGP can incorporate extrinsic evidence, including RNA-seq, IsoSeq, and cDNA alignments, loading these into a database that allows for predictions to be parallelized and scaled. AugustusCGP can also incorporate evidence from annotation sets on one or more genome in the alignment, which can help guide comparative annotation efforts by providing strong hints on where genes are expected to be found. Making use of as many of these forms of extrinsic evidence is recommended, as it greatly improves the accuracy and specificity of the model. AugustusCGP, along with all gene finding tools, can often produce false positives. For the Mouse Genomes Project , as well as the default in the CAT pipeline[68], we find that filtering for predictions with at least 2 splices reduces the false positive rate to an acceptable level. However, this can remove interesting unspliced genes like olfactory receptors [73].

1.3 Genome Annotation Pipelines

The most commonly used annotation pipeline for individual researchers is MAKER2 [39]. MAKER2 was designed to enable researchers to annotate genome assemblies they produce and not rely on institutional pipelines, and was specifically aimed at non-model and non-vertebrate species, including prokaryotes. MAKER2 combines multiple forms of extrinsic evidence with *ab-initio* prediction, and provides a fully automated end-to-end annotation pipeline. MAKER2 can make use of multiple gene prediction tools, including GeneWise, GeneMark, SNAP and AUGUSTUS. MAKER2 also performs repeat masking of the assembly.

However, MAKER2 has drawbacks. It is technically challenging to run the pipeline, and it relies on difficult to use parallel computing paradigms. MAKER2 also requires that all extrinsic evidence exist in the form of sequences, and as such requires

that RNA-seq data undergo *de-novo* assembly before use. MAKER2 does not attempt to track orthology relationships or anything about the genes predicted, and so requires subsequent processing of the gene models to determine gene family and protein domain information.

Most vertebrate annotation sets available right now are produced by large institutional pipelines at either Ensembl (Ensembl Gene Build) or NCBI (RefSeq). These institutional efforts are important for the larger size and complexity of vertebrate genomes, but as computing power becomes cheaper the need to outsource these efforts is diminishing. Turn around for annotation by RefSeq often takes months. Both of these pipelines do track orthology relationships, and assign gene common names where applicable.

This process of tracking orthology relationships will become increasingly important as the number of assembled vertebrate genomes continues to grow. A systematic framework that can track many-to-many orthology relationships is required. Ensembl has put effort into this with the Compara browser[74].

1.3.1 CAT

CAT is a annotation pipeline we have constructed that combines a variety of parameterizations of AUGUSTUS, including comparative AUGUSTUS, with transMap projections through whole-genome progressiveCactus alignments to produce an annotation set on every genome in the Cactus alignment[68]. CAT is an attempt to synthesize together all of the possible methods of genome annotation, relying on transcript projection, transcriptome and proteome alignments, simultaneous gene finding, and single genome gene finding with full length cDNA reads. CAT leverages high quality gene sets like those produced by GENCODE on mouse and human to project annotations to other genomes, augmented with predictions that add species-specificity and detects gene family expansion and collapse.

We showed that CAT was capable of leveraging GENCODE to re-annotate the rat genome, improving on the existing RefSeq and Ensembl annotations[68]. We also applied CAT to the great apes, re-annotating the existing great ape assemblies as well as annotating the new PacBio-derived great ape assemblies. In all cases, CAT

provided the highest isoform concordance compared to the Ensembl annotations when compared to a IsoSeq dataset generated from induced pluripotent stem cells from each of the great ape species. Isoform concordance for the chimpanzee annotation of the new PacBio assembly was 82.1%, the same as it was for human GRCh38 using GENCODE V27. In testing these new annotation sets by using them to quantify RNA-seq, we found an average of 9,518 more expressed genes in the great ape species than when we used the Ensembl annotations. These annotation sets greatly improved the ability to perform cross-species RNA-seq expression estimates, with Pearson r=0.96 seen when comparing CAT annotation of chimpanzee to human, compared to r=0.69 when mapping the same RNA-seq to human directly and r=0.73 when using common gene names in Ensembl V90 to perform the cross-species comparison.

1.3.2 Non-coding annotation

Comparative annotation of non-coding genes has been considered since the explosion of comparative genomics tools in the mouse genome project era[75]. MAKER2, as well as most standard annotation tools like AUGUSTUS, are not capable of detecting non-coding transcripts due to the inherent difficulty in detecting these without the strong statistical signal that protein coding transcripts provide. Tools like transMap can leverage non-coding annotations curated in high quality genomes to transfer these annotations to new genomes. However, important spliced non-coding transcripts like lincRNAs are inherently less conserved, especially at the sequence level[76, 77, 78]. This can present challenges to using whole genome alignments to annotate such transcripts. It also can be difficult to determine whether a predicted non-coding RNA is actually expressed in a species without finding it in an extrinsic dataset. An opposite problem also exists — it is not difficult to find a non-conserved ORF in a non-coding transcript, and tools like AUGUSTUS will predict coding transcripts at non-coding loci, particularly when provided extrinsic evidence that contains expression of these loci. Tools like PhyloCSF[79] can help diagnose these regions by leveraging whole genome alignments to evaluate conservation of coding signal, but are not fool-proof and the authors recommend manual curation. The phyloCSF model is being used by GENCODE to help improve manual annotation efforts.

1.4 Discussion

Recent improvements in genome sequencing technology are dropping the price of genome assembly, and particularly high quality genome assembly. Improvements in whole genome alignment have made it possible to begin to assess a new explosion of genome assemblies, giving new insight into evolution previously not possible. In this new era of comparative genomics, comparative annotation will play a central role in helping to synthesize useful information out of the deluge of data. Considering genome annotation is essential to this new era of genomes, where the question of how we will use these new genomes must be considered.

1.4.1 The future of comparative annotation

All of the comparative annotation paradigms described above are insufficient for the annotation of true many-to-many orthology. In fact, the concept of many-to-many is inherently a simplification – by definition, 1-to-1 or 1-to-many relationships are the only true relationships, with 1-to-many existing in the case where multiple duplications happened in a specific lineage.

As the number of assembled genomes grows, it will become possible to track these relationships in new ways. New approaches must be developed to assess these relationships. Cactus provides a method for generating alignments amenable to annotating this, but currently no comparative annotation tool can approach this. Comparative AUGUSTUS, for example, still relies on a alignment file format that is referenced on a specific genome. We see the possibility of a future where these problems are side-stepped through iterative use of pipelines like CAT. Gene predictions can be performed on many genomes, perhaps picking representatives evenly spread throughout the tree of life that have high quality assemblies and a decent variety of extrinsic evidence like RNA-seq available. These predictions can then be mapped via transMap to other genomes in the alignment, and then assess, combined and collapsed in a process similar to the consensus finding step that CAT currently performs.

1.5 Addendum

There are other factors to consider when talking about whole genome alignment and comparative annotation that can be overlooked. These include assembly quality, contiguity and base level accuracy.

1.5.1 Assembly quality

One issue often overlooked in the rush to assemble new genomes is assembly quality. Common metrics such as contig N50 and scaffold N50 can obscure serious problems with genome assemblies. Misjoins are common, as well as unrealistic gaps. Problems such as false tandem duplications are widespread in current reference assemblies. Too many tools currently consider genome assemblies as ground truth, without consideration for errors that are inevitably present.

1.5.2 Base level accuracy

Of particular issue to comparative annotation efforts are indel errors. These errors vastly inflate the number of transcripts seen as frame-shifted relative to a reference genome, and cause serious problems for gene finding tools leading to premature termination of gene prediction, or the introduction of false introns to bypass a false in-frame stop. Unfortunately, modern assemblies rely heavily on long-read sequencing technologies, which at this point are all inherently noisy. Much effort has been put into reducing these errors, including the introduction of diploid assembly methods [80], which reduce the rate of indel errors in the assembly caused by collapsed heterozygosity. However, our analysis of recent diploid assemblies produced by the PacBio tool Falcon Unzip[80] suggests that indel errors are still problematic and inflate the number of frame-shifting indels, with around 1% of transcripts frame-shifted in a diploid Zebrafinch assembly[81] compared to the Illumina based assembly of the same species. Additionally, these errors appear in assemblies of haploid cell lines, suggesting that heterozygosity is not the only cause.

As part of the great ape paper, techniques were developed to improve the base-level accuracy of PacBio assemblies. The currently recommended protocol involves using

Quiver or Arrow (`https://github.com/PacificBiosciences/GenomicConsensus`) to polish the assemblies, followed by Pilon [82] to perform base-level correction. Quiver is a tool published by PacBio that uses a conditional random field model of PacBio errors to correct reads mapped to a *de-novo* assembly. PacBio recommends using the newer, simpler hidden Markov model called Arrow instead, but in our experience not many groups are doing this at this point. Pilon is a tool that takes alignments of short reads to an assembly and then uses the short reads to error correct the assembly. This is powerful because it leverages the high sequence accuracy of Illumina reads to fix base level errors, but it has drawbacks. The first is simply that Illumina reads can only map to less repetitive and duplicated portions of the genome, and so the remaining portions cannot be corrected. The second is that the Pilon model has trouble with nearby heterozygous indels, leading to a sequence that does not match any haplotype present in the source sample. This led to the development of a Freebayes [83] based indel correction method that uses variant calls against the assembly to pick a haplotype at random. This process inherently produces a monoploid assembly, or an assembly that is an arbitrary mixture of all input haplotypes. However, this result is better than one that incorrectly represents any of the input haplotypes. Approaches that generate synthetic long reads, or read set technologies like 10X Genomics Linked Reads [84, 85] combined with read-cloud alignment approaches [86], allow for approaches like Pilon to improve base level accuracy on regions not directly mappable by short read alignments.

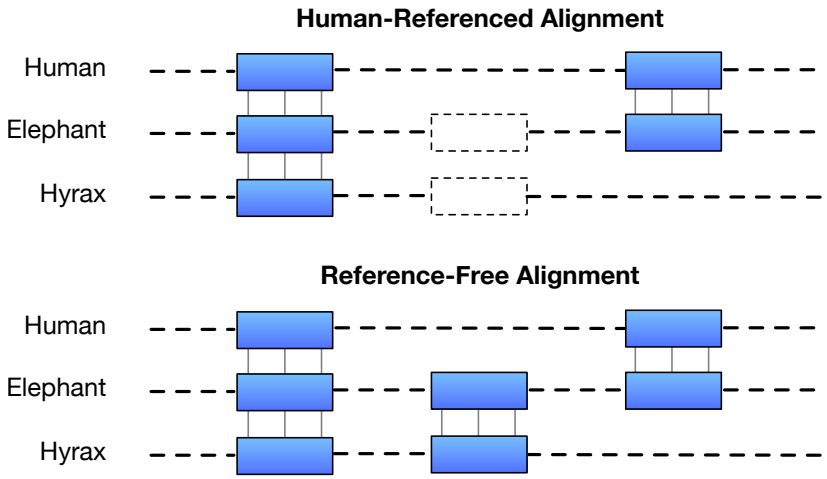


Figure 1.1: A diagram showing the difference between a reference-free and a reference-biased multiple alignment. In a human-biased multiple alignment, any regions that are deleted in human, or inserted somewhere else in the tree, cannot be aligned.

Program	Year	Description
ROSETTA	2000 [87]	Uses pairwise genomic alignments to find regions of homology. Incorporates a splice junction and exon length model.
SGP-1/-2	2001 [46]	Uses pairwise genomic alignments to find syntenic loci. Evaluates a coding and splice model in these loci.
TWINSCAN	2003 [45]	Uses local alignments between a target genome and a reference (informant) genome to identify regions of conservation.
SLAM	2003 [47]	Treats two alignments in a symmetric way, predicting pairs of transcripts.
EvoGene	2003 [88]	Phylogenetic HMM that performs *ab-initio* prediction of genes across a multiple sequence alignment (more than 2 genomes), making use of phylogenetic information.
ExoniPhy	2004 [89]	Phylogenetic HMM that performs *ab-initio* predictions across a multiple sequence alignment.
DOGFISH	2006 [90]	Two step program that combines a classifier that scores potential splice sites using a multiple sequence alignment and a *ab-initio* gene predictor that makes use of the scores from the classifier to predict gene structures.
N-SCAN	2006 [50]	Extends the TWINSCAN model to N genomes.
CONTRAST	2007 [53]	Uses a combination of SVM and CRF predictors, providing a big boost over traditional HMMs.

Table 1.1: Overview of comparative annotation tools.

Program	Year	Description
GeneWise	2004 [60]	HMM based gene prediction tool using extrinsic evidence. MAKER2 can make use of it.
NSCAN-EST	2006 [59]	HMM based gene prediction tool that makes use of EST and genomic alignments, incorporating phylogenetic information.
AUGUSTUS	2004 [62],[61]	CRF based gene prediction tool with many modes. Features are still being added. Can perform *ab-initio* gene prediction as well as incorporate extrinsic evidence. Has the ability to provide non-linear weights to various types of evidence.
EVM	2008 [91, 92]	A 'chooser' algorithm that combines previously predicted gene sets with extrinsic information to construct consensus gene sets.
PASA	2003 [93]	Uses alignments of cDNA, EST or RNA-seq to predict gene structures, including alternative splice events.
MAKER2	2008 [39, 63]	A all-in-one pipeline that runs programs including AUGUSTUS and GeneWise with extrinsic information such as RNA-seq or protein sequences to both predict annotations and construct a gene set.

Table 1.2: Overview of gene prediction tools that incorporate transcriptome data.

Program	Year	Description
Projector	2004 [70]	Similar to DOUBLESCAN, but extends the model to make use of annotation information on one sequence to inform the other. Works better than GENEWISE over long branch lengths.
AIR	2005 [71]	Integrates multiple forms of extrinsic evidence to perform alternative splice junction prediction.
transMap	2007 [55]	Uses whole genome alignments to project existing annotations from one genome to one or more other genomes.
CESAR	2016 [69]	Uses a HMM to adjust splice sites in whole genome alignments, improving transcript projections.

Table 1.3: Overview of transcript projection tools.

Program	Year	Description
MUMmer	1999 [6]	Fast aligner relying on maximal unique matches from a query sequence to a reference sequence. Recent versions remove the colinearity restriction of the first version and improve the speed.
Chains and nets	2003 [8]	Combines fragmented local alignments into larger, high-scoring "chains", which are arranged into hierarchical "nets" representing rearrangements.
Shuffle-LAGAN	2003 [94]	A "glocal" (global + local) aligner that is less restrictive than global alignment, but still enforces monotonicity of the blocks relative to one sequence.

Table 1.4: Pairwise genome alignment tools.

Program	Year	Colinear	Reference-biased	Single-copy	Description
TBA	2004 [10]	✓		✓	Multiple aligner (using MultiZ internally) that produces a collection of partially ordered "threaded blocksets."
MultiZ (autoMZ)	2004 [10]		✓	✓	Multiple alignment based on pairwise alignment from every genome to a single reference.
ABA	2004 [16]				Aligner based on the concept of A-Bruijn graphs.
EPO	2008 [15, 24]		*		Graph-based aligner allowing duplications, and optionally producing ancestral reconstructions.
VISTA-Lagan (SuperMap)	2009 [95]	*			Progressive aligner based on Shuffle-LAGAN [94].
progressive Mauve	2010 [96]		✓		Progressive aligner that attempts to remove anchors causing small rearrangements by optimizing a breakpoint-weighted score.
Cactus	2011 [27]				Graph-based aligner that attempts to remove anchors representing small rearrangements.

Table 1.5: Popular and/or historically important multiple genome alignment tools.

Chapter 2

Comparative Annotation Toolkit

This chapter goes through the history of the development of the Comparative Annotation Toolkit (CAT), and then follows with a detailed description of the CAT pipeline derived from the supplementary text of the CAT publication.

2.1 Introduction

CAT development first started in January 2015 as I took on the project of an outgoing Haussler lab member Dent Earl. Dent and Benedict were involved in the Mouse Genomes Project, an ongoing effort to assemble the genomes of sixteen laboratory mouse strains. Building on Dent's work [97], I constructed the first version of CAT, which was called comparativeAnnotator, located on GitHub at `https://github.com/ucsc-mus-strain-cactus/comparativeAnnotator`.

2.2 comparativeAnnotator

comparativeAnnotator2.2 was designed to build on transMap [55]. In developing comparativeAnnotator, I introduced the concept of classifying transcript projections based on a series of binary classifiers. These classifiers evaluate a variety of features on a transcript projection evaluating transcript completeness, identity, intron-exon structure, in-frame stops, mapping over assembly gaps, mapping to multiple contigs, and likelihood of orthology. In collaboration with Mario Stanke, building on previous work[55], compar-

ativeAnnotator also introduced the concept of AugustusTMR. This process helps clean up transcript projections, rescuing missing exons and resolving evolutionary changes to splice sites. The original version performed only AugustusTM, relying on the splice information of transMap projections and sequence composition itself without considering extrinsic evidence. After evaluating the results, and realizing that we had strain-specific RNA-seq for all strains, I worked with Mario to construct the AugustusTMR parameterization by tweaking the weights on the model to be slightly more permissive at disregarding transMap evidence in favor of RNA-seq evidence. comparativeAnnotator introduced the first version of the consensus algorithm. The consensus algorithm takes in all transcript projections for a given source transcript, and based on a heuristic combination of binary classifiers as well as alignment metrics, chooses the best orthologous candidate.

During the mouse genomes project, simultaneously to the development of comparativeAnnotator, Stefanie Nachtweide in Mario's lab was developing AugustusCGP, a novel parameterization of AUGUSTUS that performs simultaneous prediction of coding genes using a whole genome alignment. I developed an extension to the consensus algorithm that incorporates these predictions. First, I define relationships with the lift-over annotation set by comparing both exonic and intronic overlap. Then, for predictions that overlap comparative annotations, I include those that incorporate novel splice junctions supported by extrinsic evidence. Additionally, the first concept of AugustusPB was introduced during this project after IsoSeq data were generated for a few wild-derived lab strains. The version of the pipeline for that project was not fully integrated. It was not possible at the time to train AugustusCGP in an automated fashion, and not yet wrapped in Toil, and so I was not able to run it myself. All of the code to reproduce the version of the gene sets for that project that eventually got included in Ensembl release 86 are available on GitHub at `https://github.com/ucsc-mus-strain-cactus/MouseGenomesAnnotationPipeline`.

2.3 CAT development

Building on comparativeAnnotator, I worked to create a comprehensive annotation pipeline named CAT2.1. CAT integrates AugustusCGP and automates model

training, incorporates AugustusPB and wraps all of this in a Luigi and Toil based pipeline that automates genome annotation, producing both assembly hubs and annotation sets in the final product2.3.

2.4 CAT detailed description

Below is a walkthrough of the CAT pipeline. Please see the README on github (`https://github.com/ComparativeGenomicsToolkit/Comparative-Annotation-Toolkit`) for the most up to date information as well as practical information on how to run the pipeline. The text below is Supplementary Note 1 from the CAT publication.

2.5 Whole genome alignment

CAT relies on a reference-free whole genome alignment produced by the tool progressiveCactus. One or more of the genomes in the alignment should be a high quality reference whose existing annotations will be projected. Care should be taken when generating cactus alignments to provide sufficient outgroup genomes. Having high quality outgroups improves the resolution of paralogies and rearrangements.

2.6 Alignment chaining

CAT converts HAL format alignments into pairwise genome alignments via a conversion to the UCSC chain format [98]. This is accomplished by using the halLiftover tool to provide a PSL-format alignment describing each pairwise relationship to the high quality reference, and this alignment is then chained via the axtChain tool.

2.7 transMap

transMap [55, 99] is a process for using pairwise whole genome alignments to project transcript annotations from one genome to another. The main program in the Kent repository for this process is pslMap. Custom software was written for CAT and included in the Kent repository, including pslMapPostChain which chains together

mapped over transcript projection, and transMapPslToGenePred which converts the transcript projections to a gene model, keeping track of frame information and optionally filling in coding and non-coding gaps. Frameshifting gaps are not filled. CAT currently hard-codes those values at 50bp and 80bp respectively.

2.8 transMap filtering and paralogous alignment resolution

After transMap projection, alignments are filtered to their most likely ortholog, and paralogies are detected. This is performed using the tool pslCDnaFilter from the Kent repository in two parameterizations. The first parameterization does not actually filter the alignment set, but detects paralogies by relying on the localNearBest algorithm. This algorithm filters alignments based on windows of the input sequence. This process will keep multiple alignments for a given input sequence if they are in non-overlapping portions of the source transcript. This algorithm was originally designed for highly discontiguous assemblies. CAT leverages this concept to instead detect paralogies by looking at the difference. Alignments that localNearBest filters out are likely to be paralogous alignments and not instances of discontinuity, and so can be flagged as such. Putatively paralogous alignments are filtered further setting minSpan to 0.2 and providing a user-changeable minimum paralog coverage, which defaults to 50. minSpan is an effective filter against retroposed pseudogenes by filtering out any projection whose genomic size is smaller than 20% of the source transcript. The minimum paralog coverage flag does not consider any alignments whose query coverage is smaller than that value as a paralog, providing extra filtering for discontinuity. The best localNearBest parameter depends on the phylogenetic distance and assembly qualities involved, and can require tuning. The default value in CAT is 0.2, which is a fairly relaxed value. Decreasing this value will increase the rate at which discontinuity is called as paralogy.

The actual filtering of transMap projections is performed by the globalNearBest algorithm parameterized to 0, which forces pslCDnaFilter to pick the one highest scoring alignment for each input sequence. In this mode, the same minSpan value of 0.2 is used, and a minimum coverage of 10% is required. After this step, a locus resolution

step is performed to make sure that all transcripts from a given source gene end up in the same location. If globalNearBest ended up choosing multiple disjoint loci for a given gene, the highest average score locus is chosen and then lower scoring alignments for transcripts assigned elsewhere are chosen in this locus, if they exist.

2.9 AUGUSTUS

CAT runs the gene finding tool AUGUSTUS in up to four distinct parameterizations – AugustusTM (TM), AugustusTMR (TMR), AugustusCGP (CGP) and AugsutusPB (PB). The output of each of these modes is combined with the original transMap output in the consensus gene set finding process. The first two modes, TM/TMR are intended to reproduce the input isoform exactly, fixing regions where the alignment dropped an exon or introduced a small gap, or where the splice site may have shifted. These modes cannot detect novel genes or transcripts. In contrast, CGP/PB both can detect novel isoforms and genes. However, CGP can only detect one isoform for a locus and cannot find UTRs. CGP also cannot find genes in regions that did not align.

2.10 AugustusTM/AugustusTMR

The primary parameterization of AUGUSTUS for comparative annotation is primarily a method to clean up transMap projections. Due to a combination of assembly error, alignment noise and real biological changes transMap projections have frame shifting indels, missing or incomplete exons, and invalid splice sites. TM is given every protein coding transMap projection one at a time with some flanking sequence and asked to construct a transcript that closely matches the intron-exon structure that transMap provides. Since AUGUSTUS enforces a standard gene model, frame shifts and invalid splices will be adjusted to a valid form. In some cases this will mangle the transcript, producing either another isoform or something that does not resemble the source transcript. TMR runs the same inputs to AUGUSTUS, but with less strict weights on the transMap hints such that extrinsic hints from RNA-seq or IsoSeq have more bearing on the outcome. This is particularly useful in regions where an exon was

dropped in the Cactus alignment, or where a rearragement broke the alignment chains.

2.11 AugustusCGP

As TM/TMR is built on the transMap projections, it can neither identify novel genes nor existing genes of the reference annotation for which the mapping entirely failed. For this purpose, AUGUSTUS is run in its new comparative mode (CGP) recently published [72]. This mode uses a novel objective function to simultaneously predict coding transcripts in every genome in a Cactus alignment, taking in extrinsic information from any provided existing annotations as well as RNA-seq and/or IsoSeq data in any of the aligned genomes. The genome alignment is used to exploit evolutionary content for gene finding (e.g. sequence conservation, conservation of exon boundaries and selective pressure) and to transfer extrinsic evidence across genomes. The latter has the effect that each genome can benefit from the combined evidence for the clade. CGP performs best when high quality RNA-seq derived from polyA selected libraries is provided for as many genomes as possible. If this is not available, consider providing a FASTA file with previously annotated proteins of one of the currently annotated genomes.

2.12 AugustusPB

PB is run when IsoSeq data are provided and the appropriate flags set. PB runs AUGUSTUS in single genome *ab-initio* + evidence-based gene finding mode, providing high weight to extrinsic hints derived from IsoSeq data, and with the model parameterized to allow for alternative isoforms. PB provides the advantage of being able to detect genes in regions that did not align to any of the other genomes.

2.13 Parent gene assignment

CGP/PB transcripts are then assigned a possible source transcript by comparing their genomic overlap with both filtered and unfiltered transMap projections. If a transcript is assigned to an orthologous projection, then it will be evaluated for being

a novel isoform during consensus finding. If a transcript is assigned to a projection that was filtered out during paralog resolution, then it is a candidate being a possible paralog. A likely cause of this situation is a gene family expansion. If a transcript does not overlap with any transMap projections, then it is a candidate novel gene. However, the false positive rate of these is inherently high due to the likelihood of novel genes being dwarfed by the likelihood of assembly or alignment errors leading to no transMap projections in the region.

2.14 Transcript classification

transMap projections are classified by a series of classifiers that evaluate their strength. These classifiers include evaluating whether the projection was complete (100% coverage), alignment identity, whether the projection ran off the edge of a contig, whether the projection had a 1-1 ortholog relationship, and importantly how many of the exon junctions lie nearby in transcript coordinate space. This original intron classification is very important when assigning isoform relationships. Due to alignment errors and real biological changes, transMap projections may have gaps that are not near the source transcript exon junctions. The number of original introns is an important feature in the consensus finding process, protecting from retroposed pseudogenes as well as isoform switching.

Transcripts produced by TM/TMR are also classified. To do so, they are first aligned in transcript space using BLAT [100]. Alignments are performed twice, once on a whole transcript mRNA level and once using the in-frame CDS sequence using a mode of BLAT that does translation alignment. The mRNA alignments are used to perform the same original intron analysis described above, as well as record standard alignment metrics such as coverage and identity. The CDS alignments are used to evaluate transcripts for having frame shifting indels. A track of the frame shifting indels are added to the assembly hubs produced.

2.15 homGeneMapping

homGeneMapping is a tool in the Augustus package for cross-species evaluation of gene sets. It uses Cactus alignments to project the coordinates of genomic features to other genomes. Homologous gene structures are evaluated based on their consistency across species and their agreement with the combined extrinsic evidence for the clade. The latter effectively means, that a gene structure of a species with no native evidence can be "confirmed" with evidence for another species by mapping it through the genome alignment. CAT uses homGeneMapping to evaluate intron and exon features in the target genomes for 1.) consistency with the reference annotation and 2.) having extrinsic support by the combined RNA-seq and/or IsoSeq. These measures of support are used in the consensus finding process.

2.16 Consensus finding

The consensus finding process takes in transcripts from every mode and attempts to find the highest quality ortholog for a source transcript. The modes that are capable of predicting new transcripts are also evaluated for providing novel isoforms or novel loci. The final gene set is output with a series of features measuring how confident the prediction is.

To evaluate transMap, TM and TMR transcripts a consensus score is assigned to each. This score is the sum of the alignment identity target alignment coverage, intron/exon annotation support, original intron support, and intron/exon RNA-seq/IsoSeq support if extrinsic data were provided.

If CGP and/or PB is run, then the those transcripts are evaluated for providing novel information. If a prediction did not overlap any transMap projections, then it is tagged as putative novel and incorporated into the gene set. If a prediction overlaps a transMap projection that was filtered out during paralog resolution, then it is tagged as a possible paralog as well as with the names of overlapping transcripts and incorporated into the gene set. If a prediction overlaps a transMap projection and contains a splice junction not seen in the reference annotation, then it is tagged as a novel isoform and incorporated into the gene set as a member of the gene it overlapped with.

After consensus finding is complete, a final filtering process is performed. This filtering process deduplicates and strand resolves the transcript set. Duplicates most often occur when the AUGUSTUS execution modes create an identical transcript model from different input isoforms. In this case, the duplicates are removed and the remaining transcript tagged with the names of alternative source transcripts. Finally, strand resolution throws out transcripts that are on opposite strands. The correct strand is chosen by looking at which contains the most high quality transcripts.

The consensus finding process provides many user-tunable flags that can be adjusted based on the phylogenetic distances being considered. Users can change how many exons and introns should be supported by the reference annotation and extrinsic sources before being considered. Users can also decide if they want only to consider extrinsic data within that individual species or within all species in the alignment.

Another consideration is the quality of the input extrinsic data. Low quality RNA-seq data, or RNA-seq libraries not polyA selected, lead to a higher false positive rate in CGP. These often manifest as small single exon transcripts that can inflate the rate of putative novel gene calls. Adjusting a cutoff for the number of exons required to be considered novel can help drill down to a set of interesting candidates.

2.17 assemblyHubs

CAT automatically produces UCSC assemblyHubs [101], which provide a way to view assemblies, alignments and annotations on the UCSC browser without UCSC actually hosting any files. An assembly hub needs to be placed in a HTTP accessible location on the internet, and then can be loaded and viewed on the UCSC browser directly. The assembly hubs produced by CAT provide the progressiveCactus alignments as a special type of track called a snake track. These tracks allow for rearrangements to be viewed, and can be used to click through to other genomes and view the alignment from that perspective. Additionally, the reference annotations are loaded onto the reference genome, and all of the steps of CAT annotation are loaded. Splice junctions inferred from input RNA-seq files are placed in their own track. All annotation tracks have extra features added to them based on the CAT database that can be viewed by right clicking a specific transcript and looking at the details page.

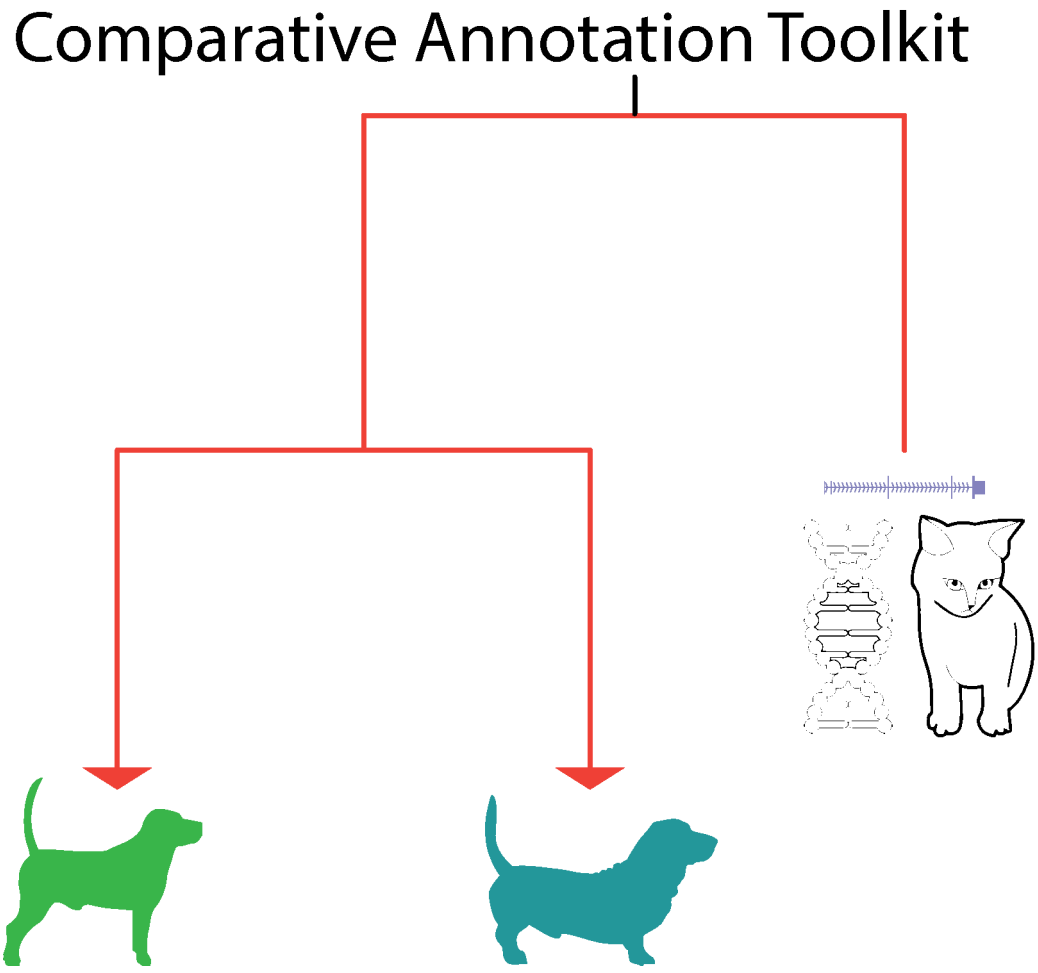


Figure 2.1: CAT pipeline logo

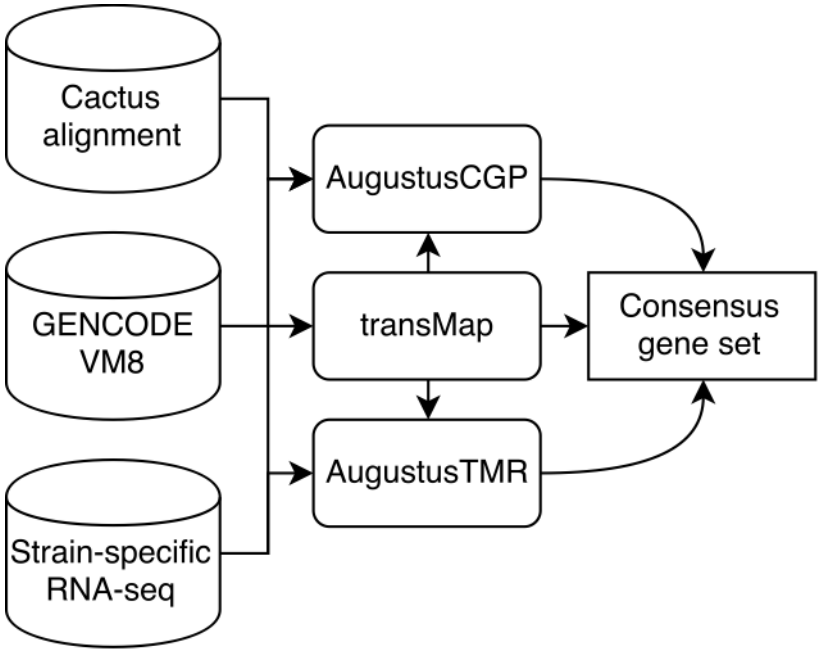


Figure 2.2: comparativeAnnotator pipeline schematic

Used in the mouse genomes paper.

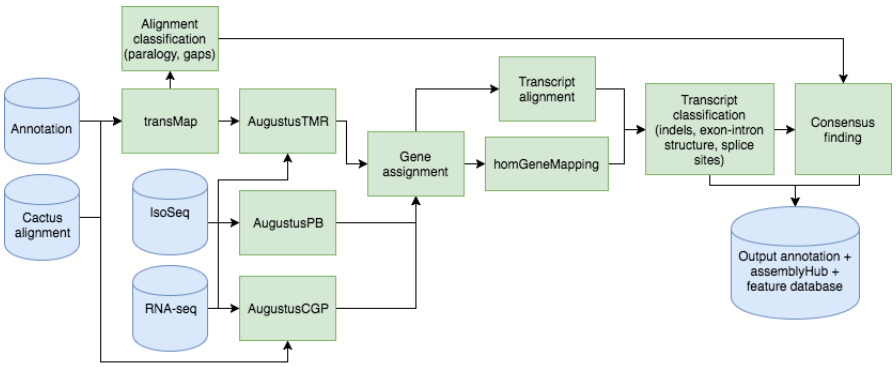


Figure 2.3: CAT pipeline schematic

CAT pipeline schematic. See the CAT pipeline description for more details.

Chapter 3

Mouse Genomes Project

This chapter introduces the Mouse Genomes Project, and is largely derived from my advancement proposal. This section goes into the history of the project, why it is important to assemble lab mouse strains, and how this project led to the development of CAT. Key results from the Mouse Genomes Project paper are highlighted. All of the figures from that paper included here are my own work except for the evolutionary history overview of Efcab3-like.

3.1 Introduction

As outlined in Chapter 2, the Mouse Genomes Project was the background for the introduction of CAT. Below, I have adapted the introduction of my PhD advancement proposal in which I explain the scientific importance of this project.

3.2 History of Inbred Mice

Laboratory mouse strains have been vital to biological, biomedical, and biochemical research since the early 20th century. The first published work involving laboratory mice came in 1902 from Lucien Cuenot, a French scientist who showed that coat colors follow Mendelian inheritance. Going further, in 1905, he showed that the A^y allele of the agouti locus, which controls coat color banding, was a lethal allele. In 1910, W.E. Castle and C.C. Little completed this experiment by showing that the Mendelian

ratios found by Cuenot were consistent when *in-utero* death of homozygous mice was accounted for.

In 1909 E.E. Tyzzer published a paper on the inheritance of tumor transplantation resistance in mice which suggested non-Mendelian inheritance of resistance. Little disagreed, and in 1914 published a theory paper suggesting that tumor resistance inheritance was instead a polygenic Mendelian trait. In 1916, Tyzzer and Little demonstrated evidence of inheritance of tumor resistance with a larger sample size. In the process of these experiments, Little began the first inbreeding crosses to create a genetically uniform strain of mice. These first crosses formed the beginning of the inbred mouse strain now called DBA [102]. Little went on to create the C57/C58 strain, and founded the Jackson Laboratory. The Jackson Laboratory today is the central repository of mouse strains and the home of mouse genetics.

The mouse strains used to found these first colonies came from the collection of Abbie Lathrop, a mouse fancier in Massachusetts. Mouse fanciers in the late 19th century collected and bred mice, crossing many wild-derived mouse strains to get desirable traits. As a result, the genetic background of most laboratory mouse strains are a mix of *Mus musculus musculus* (Eastern Europe), *Mus musculus domesticus* (Western Europe), *Mus musculus castaneus* (Southeast Asia), and the hybrid *Mus musculus molossinus* (Japan). The majority of the genetic background of laboratory mice is from *Mus musculus domesticus* [103]. In mouse research, genetic background is defined in a single-gene study situation to mean "the genotypes of all other related genes over the genome that may interact with the gene of interest and potentially influence the specific phenotype" [104]. In the case of whole genome approaches, as will be discussed here, genetic background can be generalized to mean the set of genotypic differences that contribute to phenotype differences between laboratory mouse strains.

3.3 The Mouse Genome Project

The mouse genome, published in 2002, was the second mammalian genome to be fully sequenced [105]. The assembly was based entirely on mice from the C57BL/6J strain, which had since become the standard for wild-type inbred mice. This genome, along with the transgenic and knockout mice that had been produced since the early

1980s, gave scientists an unprecedented ability to relate genetic sequence to gene function. Knockout mice with a null allele for every gene were created [106].

While individuals within an inbred mouse strain are genetically almost identical, there exists variation between strains depending on genealogy and history [107]. Phenotypic differences among strains are pronounced, and dictate which strains are useful for which experiments. For example, strains that carry the $Cdh23^{ahl}$ allele, including A/J, DBA/2J, and NOD/ShiLtJ, have progressive hearing loss starting at 3 months [108]. C57BL/6NJ mice easily get addicted to narcotics and alcohol and so are used as a model for drug addiction in humans [109]. Other strains were developed for particular disease models, such as NOD/ShiLtJ for type 1 diabetes [110] and NZO/HILtJ for type 2 diabetes [111].

Despite the wide use of various mouse phenotypes for specific research aims, very few of these phenotypes have specific molecular characterization. Many phenotypes are polygenic, and arise not just from single nucleotide polymorphisms (SNPs) but also from a complex mixture of genetic architectural changes that include small insertions and deletions, large scale rearrangements, and mutations caused by the action of retroelements [112]. These phenotypes have often been mapped to one or more quantitative trait loci (QTL), with a handful of functional variant candidates. QTLs generally have small effect sizes – that is, the percent of a quantitative trait which can be explained by the loci. Analysis of 843 QTLs in mice found an average effect size of 3% [113]. In order to better understand the genetic differences between mouse inbred strains, and how these differences lead to phenotypic differences, two projects were launched: The Mouse Genomes Project (MGP) [114] and The Mouse Phenome Database (MPD) [115].

3.4 The Mouse Genomes Project

The first genome-wide analysis of multiple mouse lines came in 2007 with microarray analysis of of 8.27 million SNPs in fifteen common mouse inbred strains [116]. These strains were chosen to cover three groups: eight classical strains used for research, three classical strains used as models for complex human diseases, and four outgroup strains derived from wild mouse populations. Wild derived strains were

included to improve understanding of the ancestral history of laboratory mouse strains.

This project was followed by phase 1 of the mouse genomes project, which sequenced a total of seventeen inbred strains with coverages ranging from 17x to 44x. Analysis of these data netted a total of 56.7 million SNPs, 8.8 million indels and 0.28 million structural variants [117]. Functional analysis found 120,000 non-synonymous substitutions and 2,051 premature stop codons. Functional variants previously characterized in a strain were found in strains not previously known to carry the variant. They found considerable phylogenetic discordance between the wild-derived strains, with three different phylogenetic histories supported by the same number of loci.

De novo assembly of these strains found 424Mb of new sequence (in contigs >100bp) not present in the C57BL/6J-derived reference genome. The variants described above represent, on average, 83.8% of the reference genome and 94.7% of the coding sequence. The remaining portions are inaccessible to short-read sequencing technologies due to low-quality mapping. For individual genomes, between 13% and 23% was inaccessible. As expected, a higher rate of inaccessibility was found in wild-derived strains due to divergence from the reference.

A genome only being 75% to 85% accessible is normal for mammals. This number is actually smaller than the true value, because genome assemblies do not fully represent repeat structures such as centromeres and telomeres. The sequence that is represented in an assembly but remains inaccessible is generally from regions of recent duplications or containing repetitive sequence which leads to multiple mappings. Structural variant breakpoint regions will also be inaccessible due to reads not mapping correctly. In order to investigate sequence in these regions, and analyze large structural variants, high quality genome assemblies of new individuals (humans) or strains (mice) need to be generated.

3.5 The Mouse Genomes Project – Phase 2

The second phase of the mouse genomes project is ongoing, and was initiated to address the shortcomings of whole genome sequencing (WGS) alignment detailed above. This phase will generate reference-quality assembly of eighteen mouse strains (Table 3.2). These include twelve inbred laboratory strains, four wild-derived inbred

strains, and two wild strains (Figure 3.1). In order to accomplish this, large insert size mate pair libraries were constructed and combined with the short insert size data used for phase 1. The assemblies were constructed and scaffolded by String Graph Assembler (SGA) [118]. Scaffolding was performed progressively by adding increasing size mate pair libraries starting at 3kb, then 6kb and 10kb. Available bacterial artificial chromosome (BAC) end sequencing data were included for the closely related laboratory strains. Optical mapping [119] and Pacific Biosciences (PacBio) long read data were included for the wild *Mus* species *Mus castaneus* and *Mus spretus*. These assemblies were tweaked and then pseudochromosomes produced.

Pseudochromosomes refer to the creation of chromosome-sized scaffolds that use information from a high-quality reference assembly to combine scaffolds and contigs from new assemblies into chromosome sized pieces. Currently, two techniques are being developed to do this.

The first tool, called Ragout (Reference-Assisted Genome Ordering UTility) [120] uses whole genome alignments to order contigs and scaffolds taking into account a phylogenetic tree of the genomes being assembled. To do this, whole genome alignments are produced between all strains and the reference using the multiple genome alignment tool progressiveCactus [121]. The second tool, being developed at Oxford by Leo Goodstadt, uses more traditional BLAST alignments to against the reference, in this case the strain C57BL6/J represented as the mm10 assembly, to order contigs.

3.6 Analysis of the Mouse Genomes Project

Genome assembly on this scale is unprecedented. Full reference-quality genome assembly of dozens of individuals of a species has not been done yet for any eukaryotic species, including humans. Mouse inbred strains provide two important advantages for undertaking such a project. The first advantage is that inbred mice are generally considered to be homozygous at every site – by 35 generations of inbreeding 0.001% of the genome will contain heterogenic tracts [122]. This is important because high-quality genome assembly of diploid individuals is difficult. Resolving variable repeat regions and structural variants can be nearly impossible without a way to phase haplotypes. Currently, phasing haplotypes is limited by the high cost of long-read sequencing or

alternative technologies such as bacterial artificial chromosome sequencing or optical mapping [119]. Standard linear representation of genomes cannot accurately represent regions of heterozygosity, particularly for complicated structural variants segregating in the population. To address this, graph-based techniques are being developed to represent all haplotypes present in a individual or individual strain.

By using inbred mice, phasing haplotypes is not necessary. The second advantage to assembly of mouse strains is that they have distinct phenotypic differences that are as of yet unexplained, but are well characterized. Mouse inbed strains are phylogenetically extremely close, having only been separated for around 50-100 years. This improves the chances of being able to correlate genotype to phenotype without being confounded by genetic noise. Mouse inbred strains can therefore be a model for using genomic information to explore phenotypic differences between individuals of a species that cannot be explained by a small number of genes.

Annotation of genomes is currently a slow and expensive process. Very few genomes currently available have a high quality gene set annotation. High quality annotation requires a combination of computational and wet-lab work to validate the annotations. More than 10 years after the first assemblies, improved annotations of both human and mouse genomes are regularly released. Starting in 2003, the Encyclopedia of DNA Elements (ENCODE) [123] project was started with the aim of a complete functional annotation of the human genome. Starting with 1% of the genome as a pilot, the project released the first full version in 2011. The Gencode Consortium [124] provides the reference gene sets to the ongoing efforts of the ENCODE project as well as other projects such as the 1000 Genomes Project [125]. In 2013, the Gencode Consortium was awarded a further grant to expand the reference gene sets to the mouse genome and continue improving the human gene sets.

The main reason that the human and mouse genome annotations are of such high quality is because of manual annotation. Through the Human and Vertebrate Analysis and Annotation (HAVANA) group at the Sanger Institute, bioinformaticians are employed to manually annotate genomes and improve genome annotation. This process is extremely slow and laborious. The HAVANA working group focuses on pseudogenes, non-coding transcripts, and splice variants, as these are more difficult for automated

annotation procedures to detect. In addition to HAVANA, the GENCODE Consortium continues to try and improve annotation of the mouse and human genome assemblies, particularly focusing on pseudogenes and noncoding elements.

3.7 Results

The annotation sets generated by this project were included in Ensembl release V86, and have been included since then. The publication for this project is pending. Figure 3.3 breaks down the number of genes comparatively annotated, while figure 3.2 breaks down the number of transcripts comparatively annotated. Figure 3.4 highlights one of the most interesting findings of the project. AugustusCGP identified a previously unannotated 188-exon gene in mouse. That this gene went unannotated for so long is a exemplar of continuing human-bias in genome annotation projects. Evolutionary analysis revealed that this long gene is conserved through vertebrate evolution, and was broken in the Hominini common ancestor by a large chromosomal inversion co-occurring with a deletion that wiped out many of the internal exons. In Hominini, both halves of the gene that remain intact appear to be expressed (Figures 3.6,3.7). The half that contains the original promoter has higher expression in GTEx data, with increased expression in testis (Figure 3.9). Fascinatingly, the other half without the promoter has measurable expression in testis data only despite a lack of histone markers suggesting that there is a functional promoter (Figure 3.8). Unfortunately, phenotyping of knockout mice did not produce a noticeable phenotype.

CAT is also useful for lifting non-coding annotations. While the majority of currently annotated lncRNAs have unknown function, some are well understood. One such example is lincRNA-cox2, which is a key regulator of inflammatory response in mice [126]. CAT was able to lift and evaluate this gene in lab mice strains, revealing a small number of single nucleotide changes present in a subset of lab mouse strains Figure (3.10).

CAT can be used in conjunction with retrogene finding tools to improve retroposed gene detection. In collaboration with the lab of Mark Gerstein, for the Mouse Genomes Project I provided annotation sets for the laboratory strains to serve as input to their pseudogene prediction pipeline. These pseudogene predictions were incorpo-

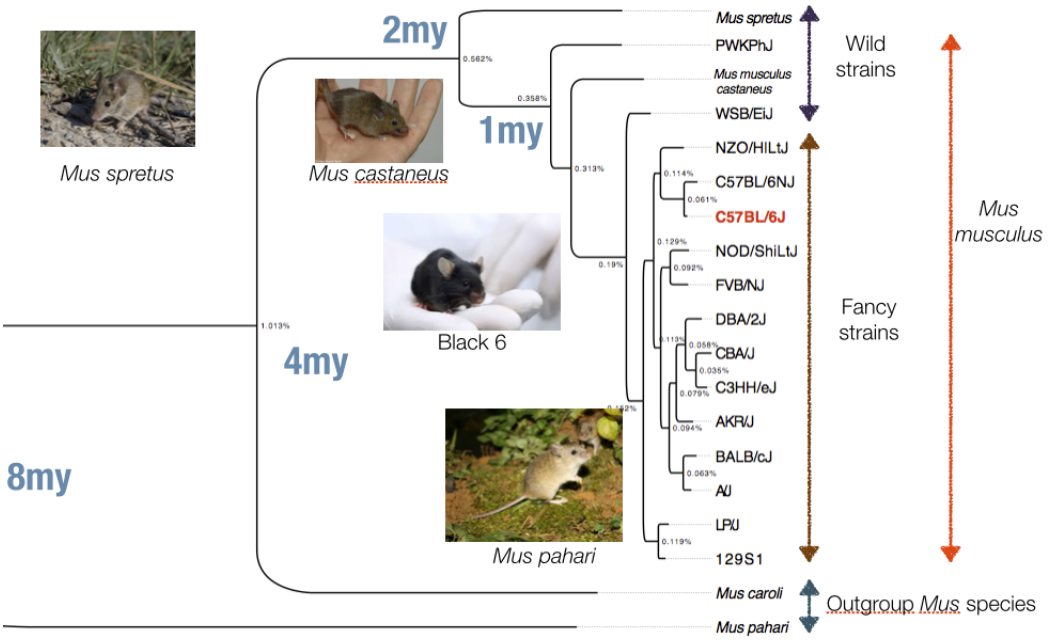


Figure 3.1: Phylogenetic tree of mouse strains in the Mouse Genomes Project Phylogenetic distances were inferred from `progressiveCactus` alignments of the assemblies and are measured as the rate of nucleotide substitutions found in the alignment. The alignment outgroup, *Rattus*, is omitted. Arrows label the groups of strains in this project. There are twelve classical inbred laboratory strains, four wild-derived inbred strains, and two wild *Mus* species.

rated into the final annotation set in combination with the direct liftover of lab mouse strains Figure (3.11).

Strain	Alleles **(unexplained)**	Description
129S1/SvImJ	$Disc1^{del}$	Developed to serve as a control inbred strain for many of the steel-derived ES cell lines (e.g. W9.5 and CJ7). Genetic background for the Knockout Mouse Project.

A/J	Ahr$^{b\text{-}2}$, **Bhr1$^{A/J}$**, **Bhr5$^{A/J}$**, Cdh23^{ahl}, Cs$^{ahl4\text{-}A/J}$, Dysf^{prmd}, Hc^{0}, Il3ra^{m1}, **Micrl^{n}**, **Naip5$^{Lgn1\text{-}s}$**, Nrg3^{ska}, **Rmcf^{S}**, Wnt9b^{clf1}, mt-Tr^{m1}	Cancer/immunology research. Progressive hearing loss between 3-5mo (Cdh23^{ahl}). Late onset progressive muscular dystrophy (Dysf). Resistant to atherosclerosis, diabetes, obesity, insulin resistance and glucose intolerance. High incidence of lung adenomas, mammary adenocarcinomas.
AKR/J	Ahr^{d}, Hc^{0}, Il3ra^{m1}, **Obq3$^{AKR/J}$**, **Obq4$^{AKR/J}$**, **Rmcf^{S}**, Soat1^{ald}	Cancer/immunology research. High leukemia incidence. Express ecotropic retrovirus AKV in all tissues. Obq alleles are unknown obesity QTLs.
BALB/cJ	**Hld**, **Micrl^{n}**	Used to generate monoclonal antibody generating plasmacytoma cell lines.
C3H/HeJ	Ahr$^{b\text{-}2}$, Gria4^{spkw1}, In(6)1J, Pde6b^{rd1}, Tir4$^{Lps\text{-}d}$	General purpose strain for cancer, infectious disease, sensorineural, and cardiovascular research.
C57BL/6NJ	Crb1^{rd8}, Cyfip2^{M1N}	Split from reference C57BL/6J in 1951. NIH strain.
CBA/J	Ahr$^{b\text{-}2}$, Pde6b^{rd1}	Blindness by wean age. Spontaneous exocrine pancreatic insufficiency syndrome.
DBA/2J	Ahr^{d}, Cdh23^{ahl}, **Fbrwt1$^{DBA/2J}$**, **Fbrwt2$^{DBA/2J}$**, Fscn2ahl8, Gpnmb^{R150X}, Hc^{0}, Klrd1$^{DBA/2J}$, Myo5^{ad}, P2rx7^{P451L}, **Tyrp1^{isa}**, A, asp2	Progressive hearing loss between 3-5mo (Cdh23^{ahl}). Progressive eye abnormalities similar to human glaucoma. Oldest inbred mouse strain.

FVB/NJ	Disc1^{del}, Hc^{0}, Pde6b^{rd1}, mt-Atp8^{m1}, Fv1	Used for transgenesis due to prominent pronuclei. Inbred for sensitivity to the Friend leukemia virus B strain. Early onset retinal degeneration (Pde6b^{rd1}). Susceptible to chemically induced tumors.
LP/J	Ahr^{d}, cdh23^{ahl}, Disc1^{del}, Ednrb^{S}	Audiogenic seizures. High incidence of late life tumors. Endothelin receptor mutation leads to white areas of coat completely lacking neural-crest derived melanocytes.
NOD/ShiLtJ	Cdh23^{ahl}, H2g7, Hc^{0}, Il^{2m1}, mt-Tr^{m1}	Polygenic model for type 1 diabetes. Insulitis, leukocytic infiltration of pancreatic islets.
NZO/HILtJ	Pctp^{R120H}	Severe obesity. Model for type 2 diabetes.
PWK/PhJ		Wild-derived. Crossed with inbred strains to create polymorphic panels. Mus musculus musculus (Prague).
SPRET/EiJ	Ahr^{b-3}	Wild-derived. Genetic mapping. Mus musculus spretus (Spain)
WSB/EiJ	A, **Kitf^{sl-blz}**	Wild-derived. Genetic mapping. Mus musculus domesticus (Maryland).
CAST/EiJ	Ahr^{d}, **Oas1b^{Flv-r}**	Wild-derived. Genetic mapping. Mus musculus castaneus (Asia).

Table 3.2: Mouse strains assembled

Phenotype information and descriptions from the Jackson Laboratory. Annotated alleles with unexplained genotypes are in boldface. The rough divergence times are included in millions of years. The branch lengths are calculated from the substitution rates seen in the `progressiveCactus` alignments of the assemblies.

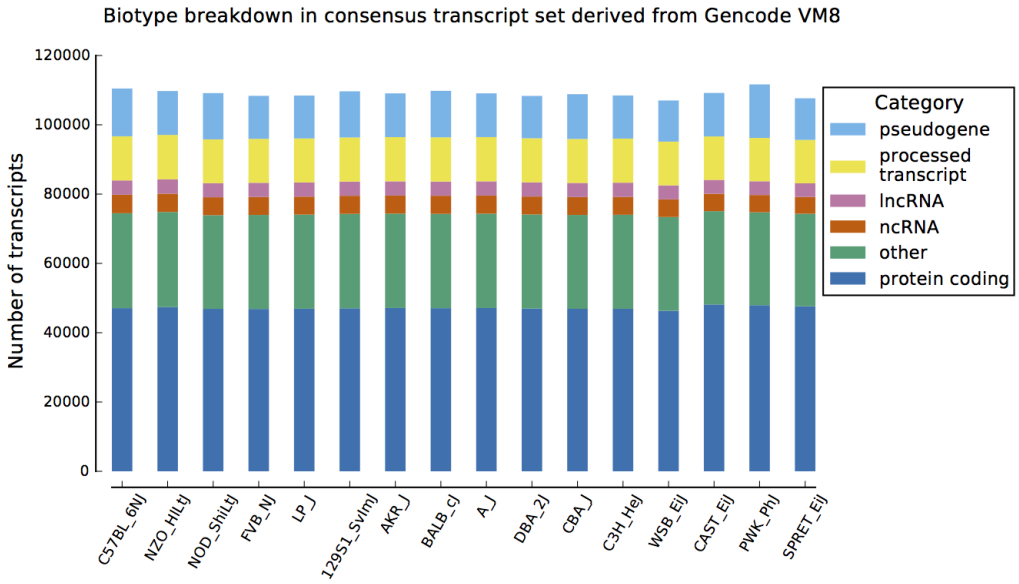


Figure 3.2: Lab mouse strain transcripts

Transcript biotype breakdown in the comparative annotation set of laboratory mouse strains.

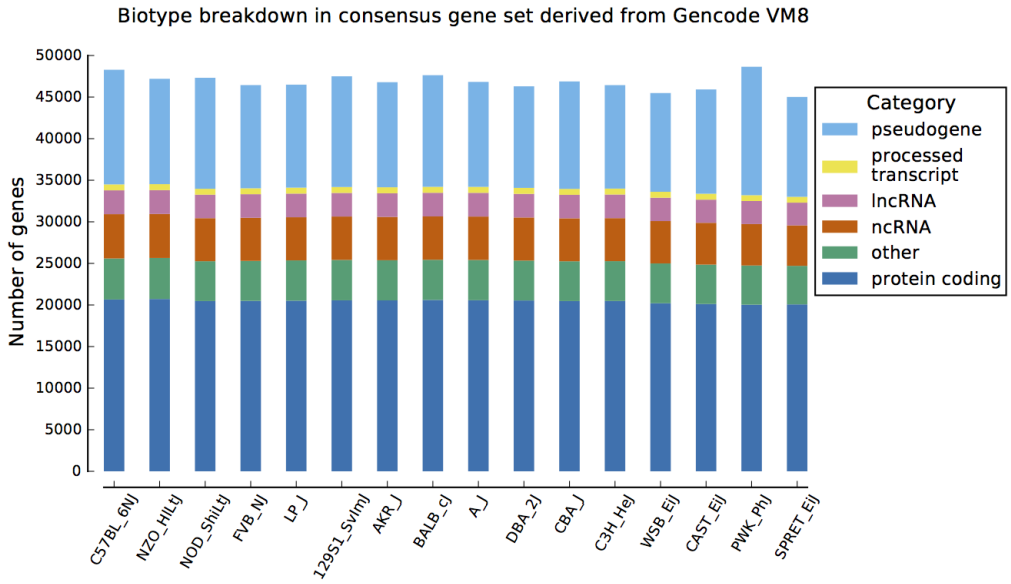


Figure 3.3: Lab mouse strain genes

Gene biotype breakdown in the comparative annotation set of laboratory mouse strains.

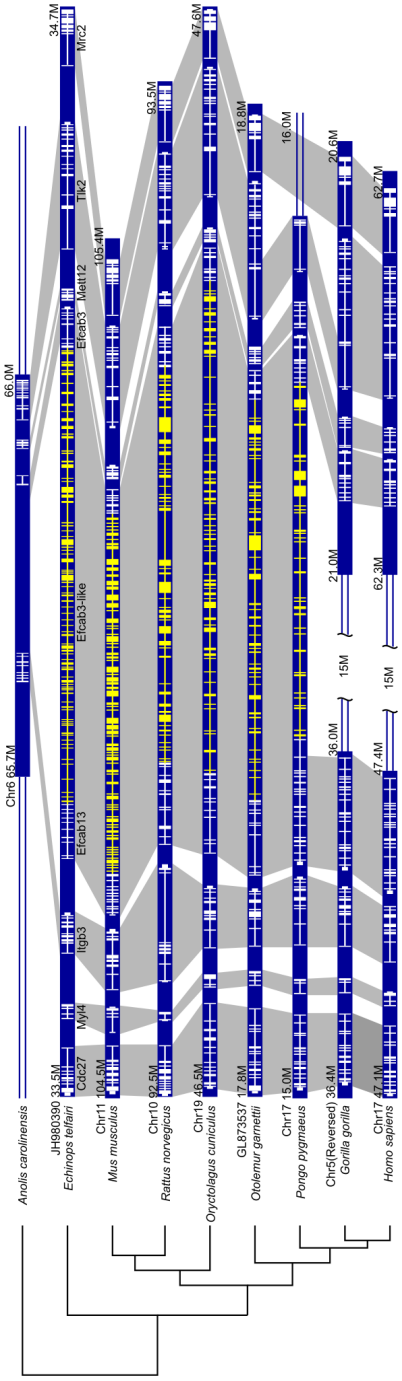


Figure 3.4: EFCAB evolutionary history

The previously unannotated 188 exon gene Efcab3-like was discovered by AugustusCGP as part of the mouse genomes project.

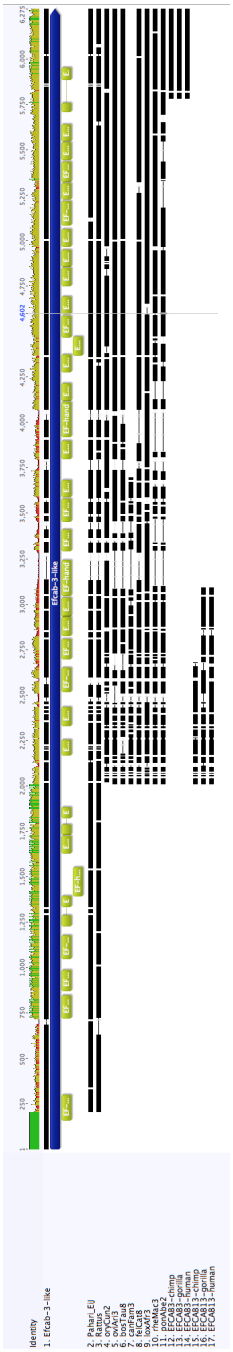
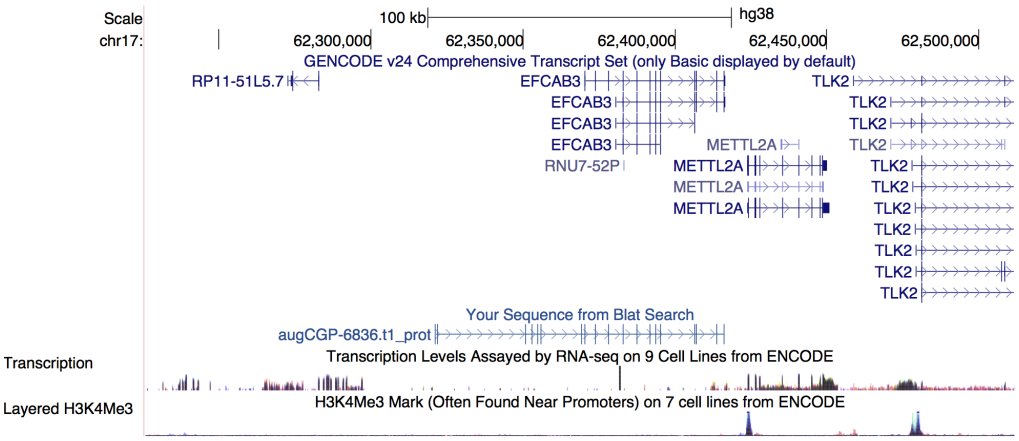


Figure 3.5: EFCAB transcript alignment

Alignment of mouse Efcab3-like to transMap projections on other mammalian genomes.

Human EFCAB3 side

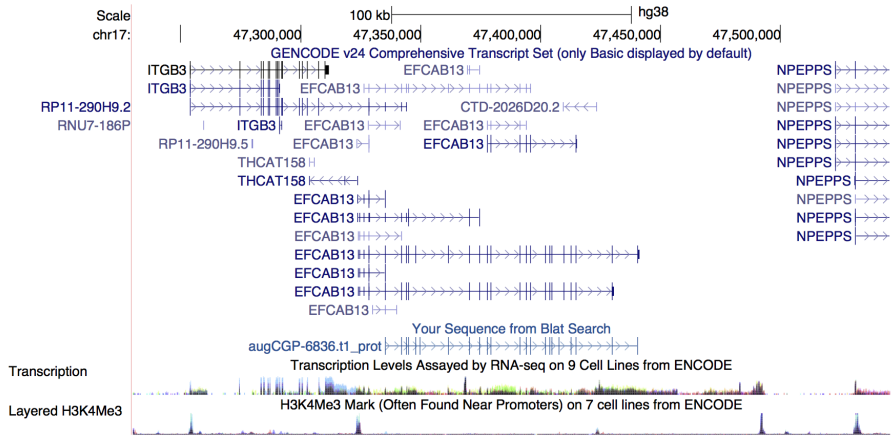


- No promoter (this side is the 3' end of the original gene)
- Minimal expression
- Annotation misses upstream exons

Figure 3.6: EFCAB3

The EFCAB3 side of the rearrangement in human has no promoter markers.

Human EFCAB13 side



- Promoter is there
- Expression

Figure 3.7: EFCAB13

The EFCAB13 side of the rearrangement in human retains the original promoter.

EFCAB3 appears to be testes-specific

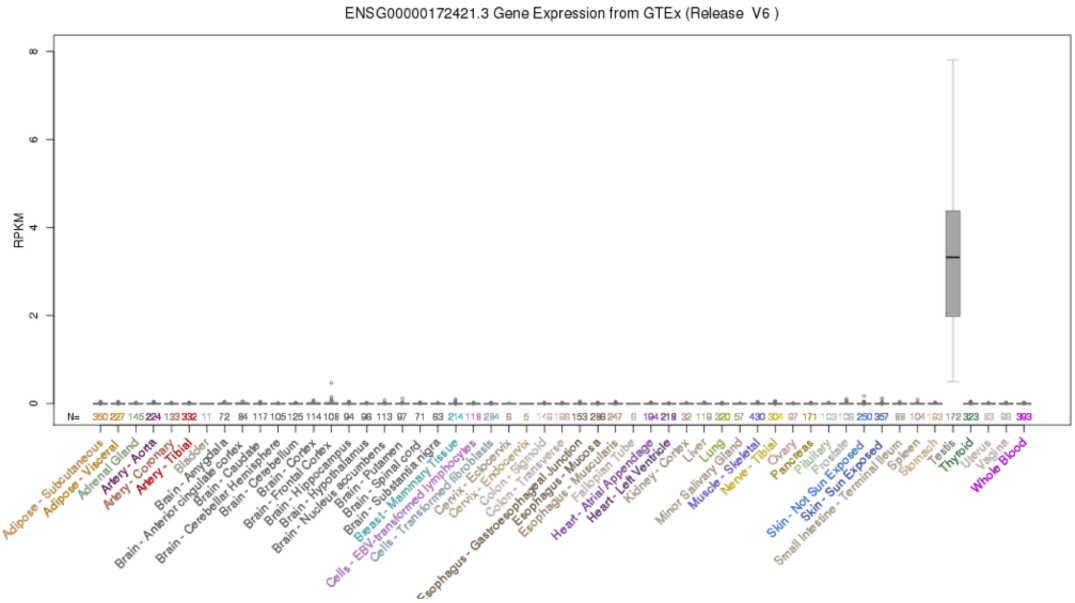


Figure 3.8: EFCAB3 expression

EFCAB3 has no measurable expression any tissue but testis, where low expression can be seen.

EFCAB13 is expressed in most tissues, with boost in testis and thyroid

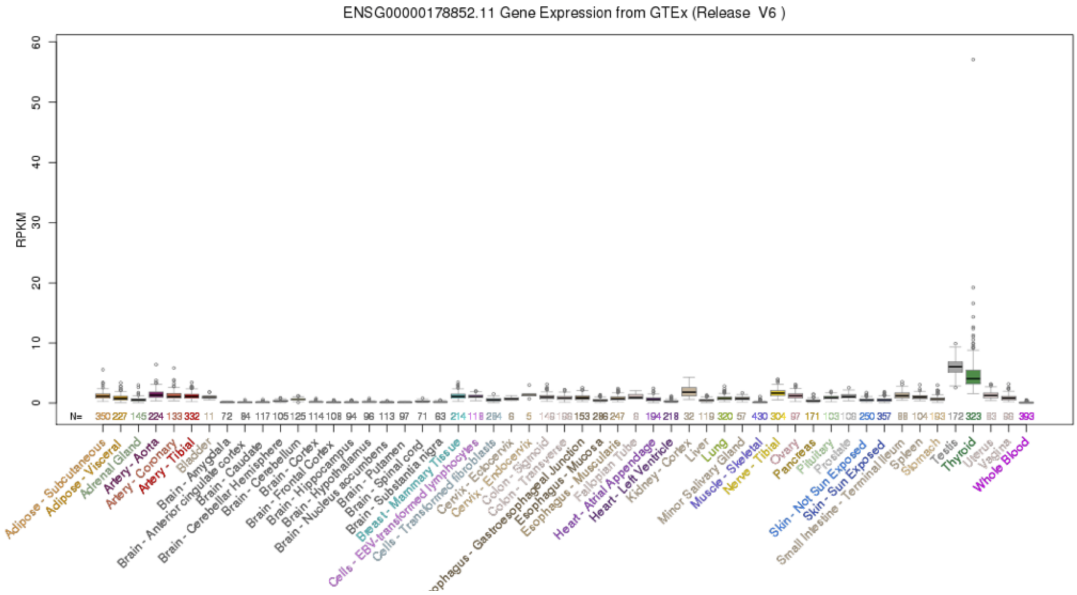


Figure 3.9: EFCAB13 expression

EFCAB13 has expression in most tissues, with a boost in testis and thyroid.

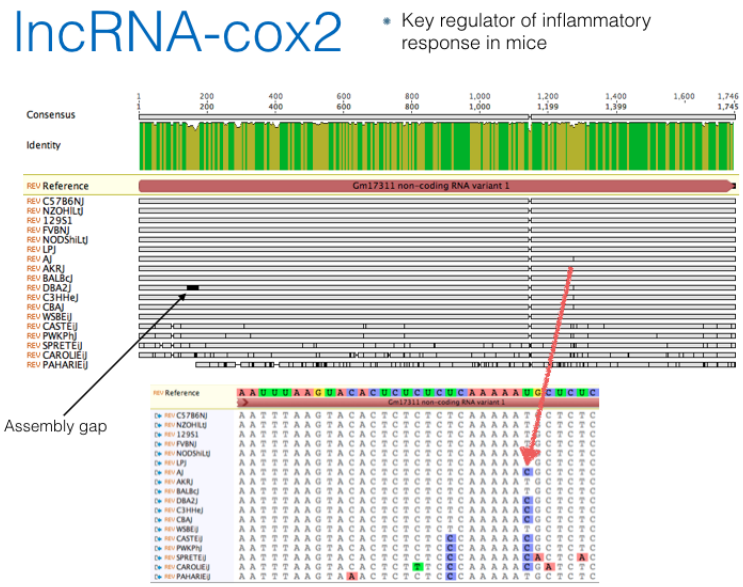


Figure 3.10: lincRNA-Cox2

lincRNA-cox2 was mapped over to lab mouse strains and a multiple sequence alignment generated of the resulting transcripts. This revealed a single nucleotide change present in a subset of lab strains. As phylogenetic distance increases in the wild-derived strains and other mus sub-species, differences accumulate.

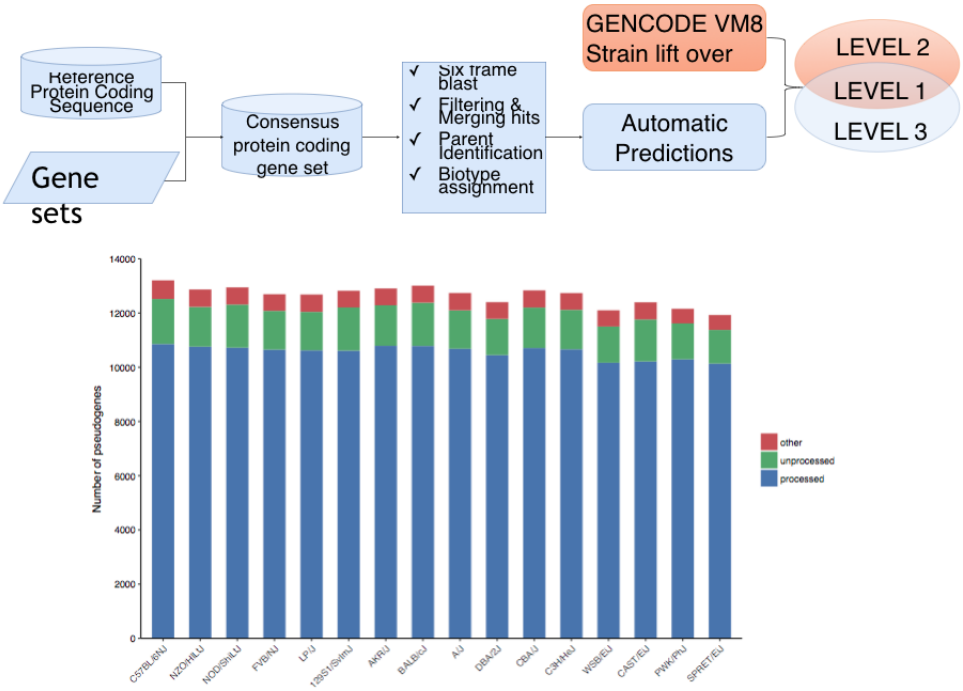


Figure 3.11: Mouse pseudogenes

The pipeline produced by Cristina Sisu that takes CAT annotations as input to evaluate pseudogenes in species (top). The number of pseudogenes predicted in each lab mouse strain, broken down by respective biotype (bottom)

	Source Institute								
	Sanger			Jackson	UNC	EBI			
Strain	Brain	B-cell	T-cell	Liver	Brain	Liver	Heart	Brain	Kidney
129S1	2			16					
AJ	2			17					
AKRJ	2								
BALBcJ	2								
C3HHeJ	2								
C57B6NJ	2								
CASTEiJ	2	5	6	16	7	2	2	2	2
CBAJ	2								
DBA2J	2								
FVBNJ									
LPJ	1								
NODShiLtJ	4			16					
NZOHlLtJ	2			16					
PWKPhJ	2			16	8				
SPRETEiJ	2								
WSBEiJ	2			16	8				
CAROLIEiJ						2	2	2	2
PAHARIEiJ						1	1	1	1

Table 3.3: RNA-seq data available for the mouse strains.

RNA-seq data are available from a variety of source institutions and tissues for each of the strains in this project. Number represent the number of biological replicates in each group.

Chapter 4

CAT paper

This chapter includes the entirety of the CAT paper, which will be submitted soon. I wrote the text of this paper and performed all of the experiments.

4.1 Introduction

We are entering a new era of genome assembly, an era in which high-quality vertebrate and plant sized genome assemblies can be created for many orders of magnitude less money than the original major reference sequencing projects.

Short read sequencing prices continue to drop and new technologies are being combined to produce assemblies of quality comparable to those previously created through intensive manual curation [127, 41, 128, 129, 40] (Kronenberg et al., submitted). These advances have allowed researchers to perform clade genomics, producing assemblies for multiple members of a species or clade [130, 131], and are required for the ambitious goals of projects such as Genome 10K [42] which aim to produce thousands of assemblies of diverse organisms. In addition, efforts are growing to produce *de-novo* assemblies of individual humans to evaluate the human health implications of structural variation and variation within regions not currently accessible with reference assisted approaches [132, 133, 134].

These advances in genome assembly require subsequent advances in genome comparison. Central to this comparison is annotation. The challenge of finding functional elements in genome assemblies has been considered for at least the past 20 years

[37]. This problem is traditionally approached by *ab initio* prediction (using statistical models of sequence composition) [62] and sequence alignment of known mRNAs or proteins[25]. The former has limited accuracy while the latter is limited by the existence of useful sequence information. Annotation pipelines such as MAKER [39], RefSeq [38] and AUGUSTUS [61] make use of both approaches. See [56] for a recent review of genome annotation methods.

A huge amount of effort has gone into the annotation of model organisms, in particular human and mouse. For the past five years, the GENCODE Consortium [30] has used a wide range of sequencing and phylogenetic information to manually build and curate comprehensive annotation sets, with over 43,281 and 60,297 ORFs in mouse and human, respectively. The GENCODE databases give a glimpse into the diversity of alternative isoforms and non-coding transcripts present in vertebrate genomes. Similarly, efforts in other model organisms such as zebrafish [135], *C. elegans* [136], *A. thaliana* [137] and many others, have produced high quality annotation sets for their respective assemblies.

As we enter a second era of genome assembly, consideration should be given to scaling annotation. Traditional genome assembly projects have been conducted by large consortia, but now individual labs can now feasibly produce high quality genome assemblies, and tools must be created to give these labs opportunities to produce high quality annotation sets on these assemblies.

Here, we present a method and toolkit to make use of multiple genome alignments produced by progressiveCactus [138] and existing high quality annotation sets to simultaneously project well curated annotations onto lesser studied genomes. In contrast to most earlier alignment methods [139, 140, 141], progressiveCactus alignments are not-reference based, include duplications, and are thus suitable for the annotation of many-to-many orthology relationships. We show how the output of this projected annotation set can be cleaned up and filtered through special application of AUGUSTUS [55], and how novel information can be introduced by combining the projected annotation set with predictions produced by Comparative Augustus [72]. These predictions can be further supplemented and validated by incorporating long range RNA-sequencing data such as those generated by the IsoSeq protocol [142]. We provide a fully featured

annotation pipeline, the Comparative Annotation Toolkit (CAT), that can perform this annotation process reproducibly on any combination of a local computer, a compute cluster, or on the cloud. We show that this pipeline can be applied to a wide range of genetic distances, from distant members of the same clade to individualized assemblies of the same species.

4.2 Results

4.2.1 Comparative Annotation Toolkit

CAT provides a software toolkit designed to perform end-to-end annotation. The only required inputs are a hierarchical alignment format (HAL) [143] multiple genome alignment as produced by progressiveCactus and a GFF3 format annotation file for the previously annotated genome(s). CAT can take as optional input a set of aligned RNA-seq or IsoSeq BAM format files, as well as protein FASTA files, which are used to construct hints for AUGUSTUS. Based on input parameters, CAT will run AUGUSTUS in up to four distinct parameterizations, two of which rely on transMap projections and two which perform *ab-initio* predictions using extrinsic information to guide prediction. The output of these modes of AUGUSTUS are evaluated alongside the original transMap projections using a combination of classifiers as well as the output from homGeneMapping [62], which uses the cactus alignments to project features such as annotations and RNA-seq support between the input genomes. A consensus finding algorithm combines these gene sets (Figure 4.1). For a more detailed description of CAT see Supplementary Note 1.

4.2.2 Annotation of great apes

The previous generation great ape assemblies (panTro4, ponAbe2 and gorGor4) as well as the new SMRT (PacBio) great ape assemblies [40] (Kronenberg et al., submitted) were annotated by CAT, using GRCh38 and GENCODE V27 as the reference. CAT identified an average of 141,477 more transcripts and 25,090 more genes in the SMRT assemblies compared to the Ensembl V91 annotation of the older great ape assemblies. Relative to the existing human annotation, the CAT annotations represent

an average of 95.0% of GENCODE gene models and 94.3% of GENCODE isoforms in the SMRT great ape assemblies. Comparing the CAT annotations of SMRT and older assemblies, we see an average increase of 610 genes (1.9%) and 3,743 isoforms (1.0%) (Supplemental Figure 4.5) in the SMRT assemblies; given this relatively small increase, most of the observed increase in genes and isoforms in the CAT annotations relative to the Ensembl annotations are therefore a result of the CAT annotation process rather than the updated assemblies.

Conversely to the overall increases in genes and isoforms, CAT identifies on average 3,553 fewer protein coding genes than Ensembl. However, this brings the total number of coding genes more closely in line with the GENCODE annotation of human, as Ensembl has an average of 2,081 more protein coding genes in great apes than GENCODE has for human (Supplemental Figure 4.5).

To evaluate these annotations in a non species-biased fashion, consensus isoform sequences created from IsoSeq reads for each species were compared to their respective species annotations. As a baseline comparison, equivalent human data was compared to the high-quality human GENCODE V27 annotation. The CAT annotation of both the SMRT and older great ape assemblies (which used the raw IsoSeq reads during the annotation process), and the Ensembl annotation of the older assemblies were compared. We calculated the rate of isoform concordance, that is the fraction of consensus IsoSeq sequences that match either exactly or fuzzily an annotated isoform (Figure 4.2A; methods). Fuzzy matching allows for the intron boundaries to shift slightly in a isoform. For the SMRT chimpanzee (74.0%/82.1% exact/fuzzy matching) and orangutan (71.4%/80.4%) genome assemblies the isoform concordance rates were comparable to the rate for human (74.6%/82.1%). The gorilla GSMRT3.2 assembly showed lower concordance (67.6%/76.9%), likely due to the higher indel error rate in that assembly (Supplemental Figure 4.6). In contrast, the isoform concordance rate for the older assemblies was lower (on average 60.0%/69.6%), mostly reflecting exons in gaps and mis-joins, and was lower still for the existing Ensembl annotations (on average 47.9%/57.6%).

To assess the utility of CAT annotations for short-read analysis of RNA expression, species-specific iPSC Illumina RNA-seq data were quantified (Figure 4.2B).

Comparing the annotations of the older assemblies, CAT identified an average of 9,518 more genes and 54,107 more transcripts with measurable expression compared to Ensembl.

We might expect the per-gene abundance estimates of the majority of genes in matched cell types to agree between species, particularly for closely related species. It is reasonable to therefore prefer *a priori* an annotation of the great apes that produces expression estimates that agree with expression estimates from the matched human data using the GENCODE annotation. Doing these comparisons, we find better correlations on average using the CAT annotation of the older assemblies (avg. Pearson r=0.63; Figure 4.2D, Supplemental Figure 4.6) than the Ensembl annotations of the older assemblies (avg. Pearson r=0.44). However, we find by far the highest correlation when CAT annotates the SMRT primate assemblies (avg. Pearson r=0.90). This reflects the increased representation in the updated assemblies of transcript sequence, especially 3' UTR regions that are important for quantifying polyA primed RNA-seq (Kronenberg et al., submitted). Notably, we find that the correlations between the CAT annotations of the SMRT assemblies and the matched human data are higher than when mapping the species specific data back to the human GENCODE annotations and comparing to the human data (4.2D, Supplemental Figure 4.6), demonstrating the benefit of having species specific annotations within closely related species that have clear cross species orthology relationships. Analysis at the isoform level showed the same patterns (Supplemental Figure 4.6), albeit with slightly weaker correlations.

Predictions performed by AugustusCGP and AugustusPB were incorporated into the gene sets based on the presence of splice junctions supported by RNA-seq or IsoSeq and not present in the transMap/AugustusTMR derived annotations (Figure 4.2C). An average of 1,677 novel isoforms and 64 novel loci were found across the assemblies with at least one IsoSeq read supporting the prediction.

CAT provides new metrics for diagnosing assembly quality. In the process of annotating the great ape genomes, we noticed that assemblies that had undergone Quiver and Pilon [82] correction still exhibited a systematic bias towards coding deletions. These were identified to be related to heterozygosity in the input dataset and a variant calling based correction method (Kronenberg et al., submitted) was developed

to resolve these issues, dramatically lowering the coding indel rate and reducing systematic bias (Supplemental Figure 4.6). CAT can also measure gene assembly contiguity by reporting the number of genes whose transcripts end up split across multiple contigs, or on disjoint intervals in the same contig. Comparison of split gene metrics between the old and new primate assemblies shows 504 fewer split genes in chimpanzee, 560 fewer in gorilla and 1,858 fewer in orangutan (Supplemental Figure 4.7).

4.2.3 Annotation of personal human diploid assemblies

High-quality de novo assembly of a human genome is increasingly feasible; both Pacific Biosciences [80, 144, 81] and 10x Genomics [145] provide tools to construct phased, diploid assemblies. Annotating diploid assemblies provides a window into haplotype-specific structural variation that may affect gene expression. To evaluate the ability of CAT to provide this analysis a progressiveCactus alignment was generated between hg38 and the two haploid cell line assemblies, CHM1 (GCA_001297185.1) and CHM13 (GCA_000983455.2), and CAT used to annotate the resulting pseudo-diploid CHM1/CHM13 genome in the alignment. 99.3% of genes present in GENCODE V27 were identified in CHM1, and 99.2% in CHM13 (Figure 4.3A).

After filtering, 733 genes in CHM1 and 585 genes in CHM13 were identified with frame-shifting indels. Compared to ExAC, which found between 75 and 125 putative truncating events per individual [146], this result suggests indel errors in the assemblies are producing false positives. Both assemblies exhibit a systematic overrepresentation of deletions despite coming from haploid cell lines (Figure 4.3B, left). Split gene analysis found the CHM1 assembly to be slightly more gene contiguous (Figure 4.3B, right).

Manual analysis of the 27 genes missing in CHM13 relative to CHM1 and the 6 genes missing in CHM13 relative to CHM1 led to the discovery of the example region in Figure 4.3C. This deletion removes most of the exons of TRIB3, a pseudokinase associated with type 2 diabetes [147] (Supplemental Figure 4.8).

4.2.4 Reannotating the rat genome

To test the ability of CAT to translate annotations between more distantly related genomes, we tested CAT's ability to reannotate the rat genome using information from the mouse genome. These genomes differ by approximately 0.18 substitutions/site, much more, for example, than the 0.04 substitutions/site separating the human and orangutan genomes[148].

CAT was run on a Cactus alignment between mouse (mm10) and rat (rn6) using rabbit (oryCun2), Egyptian jerboa (jacJac1) and human (hg38) as outgroups. To provide hints to AUGUSTUS, RNA-seq data were obtained from SRA [149, 150, 151] (Supplemental Table 4.2). For comparison we used existing Ensembl and RefSeq rat annotations and ran the MAKER2 pipeline [39] to generate an annotation set. MAKER was provided both a Trinity [92] *de-novo* assembly of the input RNA-seq data provided to CAT (MAKER does not process raw RNA-seq) as well as the mouse protein sequences from GENCODE VM11, together providing a comparable input set to what CAT had.

CAT comparatively annotated 78.1% of genes and 91.9% of protein coding genes present in GENCODE VM11 on rn6 (Supplemental Figure 4.9), representing an increase of 14,675 genes and 74,308 transcripts over Ensembl V90, 5,104 genes and 32,157 transcripts over RefSeq and 14,541 genes and 81,022 transcripts over Maker. 13,651 loci were identified with no overlap to any other annotation set (Supplemental Figure 4.10).

We compared CDS exon and CDS intron predictions between annotation sets (Supplemental Figure 4.10). Ensembl, RefSeq and CAT CDS Exon annotations were all comparably similar (between 0.648 and 0.659 Jaccard similarity), while for CDS introns CAT and RefSeq displayed the highest Jaccard similarity (0.841). In all comparisons MAKER was the outlier (Figure 4.4A) with the lowest similarity to the other sets.

The input RNA-seq dataset was used for isoform quantification against the CAT, MAKER, Ensembl and RefSeq transcriptomes (Figure 4.4B). CAT identified 1,881 protein coding genes and 1,011 lncRNAs with measurable expression not present in either Ensembl or RefSeq. CAT also identified 27,712 expressed coding splice junctions not in the union of RefSeq and Ensembl, for a total of 21,267 novel expressed isoforms. 5,526 of the 13,651 loci unique to CAT had measurable expression.

AugustusTMR, which uses TransMap and RNA-seq, provides CAT with a way to improve transcript predictions projected between species. Comparing the 10,335 multi-exon protein coding transcripts in which the AugustusTMR prediction differed from the input transMap projection, we see considerable overall improvement in resulting RNA-seq support of predicted splice boundaries in the AugustusTMR transcripts (Figure 4.4C).

4.2.5 Annotation of a diverse set of mammals

Finally to test CAT's ability to annotate across a substantial and diverse range of genomes, 13 mammalian genomes were comparatively annotated using the mouse (mm10) GENCODE VM15 as the reference transcript set (Figure 4.5A). Species-specific RNA-seq was used for every genome (Supplemental Table 4.2). To assess the completeness of these annotation sets, 4,104 benchmarking universal single copy orthologs (BUSCO) were used [152], which by design should be nearly uniformly present in each of these genomes. On average, only 108 BUSCO genes (2.63%) were not annotated by CAT in each genome (Supplementary Table 4.1).

To estimate the usefulness of these annotation sets, the input RNA-seq datasets were used to quantify expression of the annotation sets (Figure 4.5B). The main factor in measurable expression is the variety of the input RNA-seq datasets, as exemplified by the ability to measure expression of 88.9% of genes annotated in the sheep genome.

To assess the CAT translation of annotations over large phylogenetic distances, the annotation of human hg19 (GRCh37) produced in the representative mammalian genome annotation was compared to the current human GENCODE annotation set for that assembly (GENCODE V27lift37). Of the 19,233 ICE isoforms detected when running ToFu [142] against hg19, 12,911 (67.15%) fuzzy matched a CAT isoform compared to 15,920 (82.8%) of the human GENCODE annotations (Figure 4.5C). The Jaccard similarity of the human CAT annotation to the human GENCODE annotation set was similar to the similarities seen in the rat comparison (Figure 4.5D).

4.3 Discussion

Gene annotation is a longstanding and critical task in genome informatics that must now be scaled to handle the rapidly increasing number of available genomes. At the time of writing there were XX vertebrate genomes available from NCBI, but only YY (YY%) had available gene annotation sets, and only ZZ (ZZ%) and WW (WW%) had Ensembl and RefSeq annotations, respectively.

We introduce CAT to help meet this need, building around a number of key innovations. Firstly, CAT utilizes the reference-free, duplication aware multiple genome alignments we have developed. This allows CAT to annotate multiple genomes symmetrically and simultaneously, breaking from the traditional pattern of annotating each new genome individually, as is currently the practice for the RefSeq, Ensembl and MAKER gene building pipelines. Not only does this solve a key scalability issue, by annotating multiple genomes simultaneously and consistently CAT is able to produce orthology mappings, naming each equivalence class of orthologs based upon an initial reference annotation, and adding to this sets of newly discovered genes. This can provide valuable evolutionary insights. For example, the analysis of the rat genome shows that many of the alternative isoforms and projected transcription start sites identified by GENCODE in mouse genes are supported by expression analysis in rat (Supplemental Figure 4.11).

A second key innovation made by CAT is its leveraging of existing reference annotations. A huge amount of effort has been placed into the annotation of key species, such as human and mouse, employing myriad technologies and extensive, labor intensive manual curation. It is very unlikely that this effort will be replicated across a significant fraction of other genomes, so instead we propose the "project and augment" strategy employed by CAT to annotate related genomes. Here we show that this strategy is very clearly able to improve the annotation of great ape genomes, using the human GENCODE set as the reference, and we make the case that we can even improve the annotation of a genome as well studied as the rat.

To circumvent the reference bias of existing annotations and to discover new genes and isoforms CAT is able to integrate multiple forms of extrinsic information, using multiple, novel parameterizations of the AUGUSTUS algorithms. This includes use of new long-read RNA data, in particular IsoSeq data, and shortly will integrate

nanopore based long-read data[153]. Using this expression data not only allowed us to confirm expression of a substantial fraction of isoforms, but allowed us to discover thousands of novel isoforms and dozens of novel genes in the great ape genomes.

With the advent of more affordable *de novo* genome assembly there is renewed interest in the generation of *de novo* human genomes, and in general the creation of multiple *de novo* genomes for a species. This has the advantage of providing fully independent reconstruction and is particularly appropriate for sequences that are highly divergent from the reference, e.g. structural variations. However, such assemblies do not negate the need for genome comparison; Cactus can be parameterized to rapidly create sensitive whole genome alignments of human genomes, and here we have demonstrated that CAT can be used to build upon this a high quality diploid gene annotation and ortholog mapping.

CAT works best when provided RNA-seq data, but for many species this may not be possible. From our experience, a reasonable amount (on the order of 50 million reads) of RNA-seq from tissues like brain and liver is fairly informative. Using polyA selected libraries is recommended, as it greatly reduces false positive predictions in AugustusCGP. IsoSeq data allowed for the discovery of thousands of novel isoforms in the great apes, but may be too expensive for many projects. In clade genomics projects, we would suggest generating RNA-seq for a few of the species and then reliance on the coordinate mapping that AugustusCGP and homGeneMapping provide to evaluate support in other members of the clade.

A key barrier to the use of bioinformatics tools is their ease of use; we have focused on providing cloud agnostic distributions of the CAT software so that, despite its complexity, it can be run within a uniform computational environment by external groups.

CAT is not without limitations. In future it would be good to use the genome alignments to not only project transcripts, but to use the evolutionary conservation signatures to predict the potential likelihood of projected annotations being coding [79]. CAT also does not yet provide means to detect new processed, unprocessed and unitary pseudogene predictions other than via projection of existing annotations. CATs current implementation also does not attempt to put weights on the features used for

constructing a consensus gene set. Instead, it simply scores transcripts based on the sum of all features evaluated. In the future deep learning methods could be added to CAT to construct feature weights and improve consensus finding, and so better mimic the labor intensive efforts of manual annotators who currently weigh such evidence.

An earlier version of CAT was used to annotate the PacBio-based assembly of the gorilla genome [40] as well as produce the current Ensembl annotations for 16 laboratory mouse strains as part of the Mouse Genomes Project (`http://www.sanger.ac.uk/science/data/mouse-genomes-project`). In addition, CAT has been proposed for the Vertebrate Genomes Project (VGP), which aims as a pilot project to assemble and annotate one member of every order of vertebrate species. CAT also will be used on the 200 Mammals Project, which aims to add approximately 140 new mammalian genome assemblies to the existing set (`https://karlssonlab.org/2017/08/03/the-200-mammals-project/`). These projects will provide a new understanding of gene evolution.

4.4 Materials and Methods

CAT produces as output a series of diagnostic plots, an annotation set for each target genome, and a UCSC comparative assembly hub [101]. Both the pipeline and associated documentation can be found at `https://github.com/ComparativeGenomicsToolkit/Comparative-Annotation-Toolkit`. CAT is constructed using the Luigi (`https://github.com/spotify/luigi`) workflow manager, with Toil [154] used for computationally intensive steps that work best when submitted to a compute cluster.

4.4.1 RNA-seq

CAT annotation is improved when species-specific RNA-seq data are provided. These data are used as hints for AugustusTMR and AugustusCGP. In AugustusTMR, RNA-seq helps fill in missing information in the alignment, as well as resolve evolutionary changes. In AugustusCGP, RNA-seq additionally helps prevent false positives inherent in *ab-initio* gene finding. For these reasons, RNA-seq was obtained from SRA for all species annotated in this paper. All RNA-seq were aligned to their respective genomes with STAR [155] and the resulting BAM files passed to CAT to construct the

extrinsic hints database. See Table 4.2 for accessions and tissue types of RNA-seq data used for annotation. In addition, for the PacBio great ape annotation, RNA-seq data were generated using iPS cell lines for human, chimpanzee, gorilla and orangutan derived from cells for the same individuals as the assemblies (Kronenberg et al., submitted). For all expression analyses, Kallisto [156] was used.

4.4.2 Annotation set similarity analysis

Jaccard similarity analysis was performed with bedtools [157]. The rat locus overlap analysis was performed with the Kent tool clusterGenes, which requires exonic overlap on the same strand.

4.4.3 IsoSeq

IsoSeq full-length non-chimeric reads (FLNC) were also generated from the great ape iPSC lines and aligned with GMAP [158].To perform isoform level validation in the primates, the IsoSeq data used as input to CAT were also clustered with isoform-level clustering (ICE) and then collapsed into isoforms using ToFu [142]. Ensembl provided a pre-release of their new V91 annotations for panTro4 and gorGor4, but did not yet run their updated pipeline on ponAbe2.

4.4.4 ICE validation

The output transcripts from ICE were compared to various annotation sets in both an exact and fuzzy matching scheme. In the exact scheme, the genomic order and positions of all of the introns (an intron chain) of a transcript are compared to any ICE isoforms which overlap it. In the fuzzy matching scheme, each annotated intron chain is allowed to move up to 8 bases in either direction and still be called a match.

4.4.5 BUSCO

The mammalian BUSCO analysis was performed using the mammalia_odb9 set of 4,104 genes. BUSCO was ran against the complete protein coding sequence repertoire produced by CAT in that species in the 'protein' mode.

4.4.6 progressiveCactus

All cactus alignments except the 14-way mammal alignment were generated using progressiveCactus (`https://github.com/glennhickey/progressiveCactus`) commit 91d6344. For the mouse-rat alignment the guide tree was

```
(((Lesser_Egyptian_jerboa:0.1,(Mouse:0.084509,Rat:0.091589)mouse_rat:0.107923)
    rodent:0.148738,Rabbit:0.21569)glires:0.015313,Human:0.143908);
```

For the primate alignments, the guide tree was

```
(((((((Susie_Gorilla:0.008964,(Human:0.00655,Clint_Chimp:0.00684)human_chimp
    :0.00122)gorilla_chimp_human:0.009693,Susie_Orangutan:0.01894)great_ape
    :0.003471,Gibbon:0.02227)great_ape_gibbon:0.01204,Rhesus:0.004991)
    old_world_monkey:0.02183,Squirrel_monkey:0.01035)monkey:0.05209,Bushbaby
    :0.1194)primate_anc:0.013494,Mouse:0.084509);
```

An identical tree (with different assembly names) was used for the alignment of current reference great apes. For the pseudo-diploid human alignment, the guide tree was

```
((hg38:0.001,chm1:0.001,chm13:0.001)human:0.01,chimp:0.01);
```

representing a star phylogeny of the three human assemblies. For the 14-way mammal alignment, the progressiveCactus commit used was e3c6055 and the guide tree was

```
((((oryCun2:0.21,((Pahari_EiJ:0.03,mm10:0.025107)1:0.02,rn6:0.013)1:0.252)
    1:0.01,((((hg19:0.00642915,panTro4:0.00638042)1:0.00217637,gorGor3
    :0.00882142)1:0.00935116,ponAbe2:0.0185056)1:0.00440069,rheMac3:0.007)
    1:0.1)1:0.02,((oviAri3:0.019,bosTau8:0.0506)1:0.17,(canFam3:0.11,felCat8
    :0.08)1:0.06)1:0.02)1:0.02,loxAfr3:0.15);
```

Slightly out-of-date versions of some assemblies (hg19 and rheMac3) were used because a collaborator had data on those assemblies that they wished to use the alignment to analyze. The rodent and primate subtrees were first aligned separately (the rodent subtree originally included additional mouse strains). The two subtrees were then stitched together into a single alignment by aligning together their roots along with several Laurasiatheria genomes. This was done to save alignment time by reusing existing alignments.

4.4.7 CAT

CAT was ran on the UCSC Genome Browser compute cluster for all annotation efforts in this publication. CAT commit f89a814 was used. For a detailed description of how CAT works, see both the supplementary text as well as the README.md on github (`https://github.com/ComparativeGenomicsToolkit/Comparative-Annotation-Toolkit`).

4.4.8 Pipeline runtime

CAT is relatively efficient, taking on the order of thousands of core hours to run. The largest considerations for runtime are running the various parameterizations of AUGUSTUS. AugustusCGP may run significantly faster on alignments with many genomes by reducing the chunk size from the default, but at the cost of lower quality predictions. AugustusTMR runtime scales linearly with the number of protein coding genes in the input annotation set, but scales non-linearly with the number of extrinsic hints provided, particularly if the hints are contradictory. Running CAT on the PacBio primate genomes took a total of 7,030 core hours, with 3,437 of those dedicated to running AugustusTMR, 1,191 dedicated to running AugustusPB, and 2,190 dedicated to running AugustusCGP. Running CAT on the 14-way mammalian alignment took a total of 24,122 core hours, with 14,045 of those dedicated to running AugustusTMR, and 8,225 dedicated to running AugustusCGP.

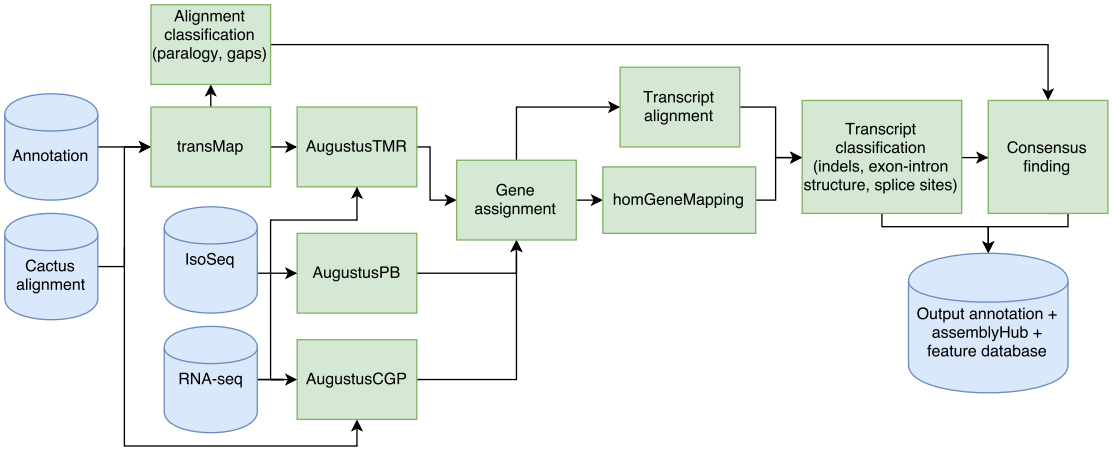


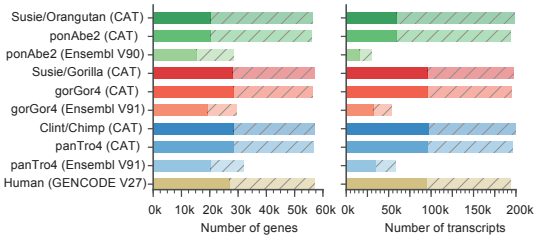
Figure 4.1: CAT pipeline schematic

The CAT pipeline takes as input a HAL alignment file, an existing annotation set and aligned RNA sequencing reads. CAT uses the Cactus alignment to project annotations to other genomes using transMap [55]. These transcript projections are then filtered and classified. Optionally, AUGUSTUS can be run into up to three parameterizations. 1) AugustusTMR, treats each transcript projection separately and fixes errors in projection. 2) AugustusPB, uses long read RNA sequencing to look for novel isoforms. 3) AugustusCGP [72] uses the Cactus alignment to simultaneously predict protein coding genes in all aligned genomes. CGP and PB predictions are assigned to transMap projections where possible. All transcripts are classified for extrinsic support and structure and a chooser algorithm picks the best representative for each input transcript, incorporating CGP and PB transcripts when they provide novel supported information. The final consensus gene set as well as associated feature tracks are used to create a assembly hub ready to be loaded by the UCSC browser.

A) Isoform level accuracy compared using full length transcript sequencing of species-specific iPSC RNA

	Human (GENCODE V27)	Clint/ Chimpanzee (CAT)	Susie/ Gorilla (CAT)	Susie/ Orangutan (CAT)	panTro4 (CAT)	gorGor4 (CAT)	ponAbe2 (CAT)	panTro4 (Ensembl V91)	gorGor4 (Ensembl V91)	ponAbe2 (Ensembl V90)
Multi-exon collapsed ICE transcripts	19,271	18,863	22,383	14,377	17,455	20,046	13,102	18,665	20,046	13,102
Exactly matches annotation	14,379 (74.6%)	13,950 (74.0%)	15,137 (67.6%)	10,258 (71.4%)	11,175 (64.0%)	12,789 (63.8%)	6,857 (52.3%)	10,252 (55.0%)	10,130 (50.5%)	5,000 (38.2%)
Fuzzy matches (±8bp) annotation	15,826 (82.1%)	15,485 (82.1%)	17,201 (76.9%)	11,563 (80.4%)	12,755 (73.1%)	14,459 (72.2%)	8,303 (63.4%)	11,995 (64.3%)	11,901 (59.4%)	6,426 (49.1%)

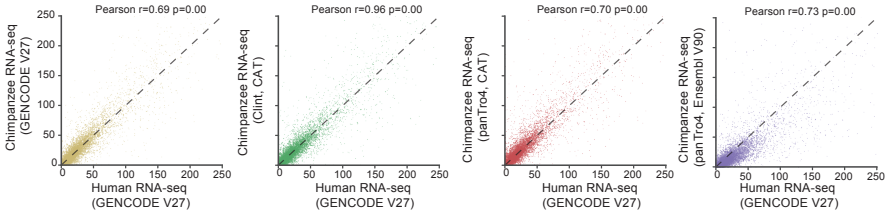
B) Non-zero expression estimates of species-specific iPSC RNA-seq (Kallisto)



C) Incorporation of AugustusCGP and AugustusPB predictions

	Clint/ Chimp	Susie/ Gorilla	Susie/ Orangutan
AugustusPB putative novel isoform	1720	1251	1425
AugustusCGP putative novel isoform	265	192	178
AugustusPB putative paralogous loci	52	48	58
AugustusCGP putative paralogous loci	14	11	9

D) Per gene TPM comparison of different chimpanzee assembly and annotations to human



Assembly/Annotation Pair	GRCh38/GENCODE V27	Clint/CAT	panTro4/CAT	panTro4/Ensembl V90
Number of coding genes with TPM > 0.1	16,098	15,277	15,242	13,463

Figure 4.2: Primate annotation

A) Validating CAT annotations using IsoSeq data. As a baseline comparison, IsoSeq data from human iPSC cells was compared to the GENCODE V27 annotation. IsoSeq data from chimpanzee, gorilla and orangutan iPSC lines were compared to respective species specific annotations. The IsoSeq data were clustered with ICE and collapsed using ToFu[142]. CAT annotation of PacBio great apes showed similar isoform concordance to human, and improvement over the older assemblies. B) Kallisto[156] was used to quantify liver Illumina RNA-seq from each species on both the gene and transcript level on the existing and new great ape assemblies. Solid bars are transcripts or genes with TPM >0.1, while shaded hatched bars are the remainder of the annotation sets. CAT annotation of great apes show nearly the same number of expressed genes and isoforms as the GENCODE reference on human with the exception of orangutan. C) The number of novel isoforms and paralogous genes with IsoSeq support discovered by analysis of AugustusPB and Augustus CGP predictions for each species. D) Kallisto protein coding gene-level expression for chimpanzee iPSC RNA-seq is compared to human across all of the chimpanzee annotation and assembly combinations as well as when mapped directly to human. In all cases the x-axis is the TPM of human iPSC data mapped to human. The highest correlation (Pearson r=0.96) is seen when comparing Clint annotated with CAT to GRCh38 annotated with GENCODE V27.

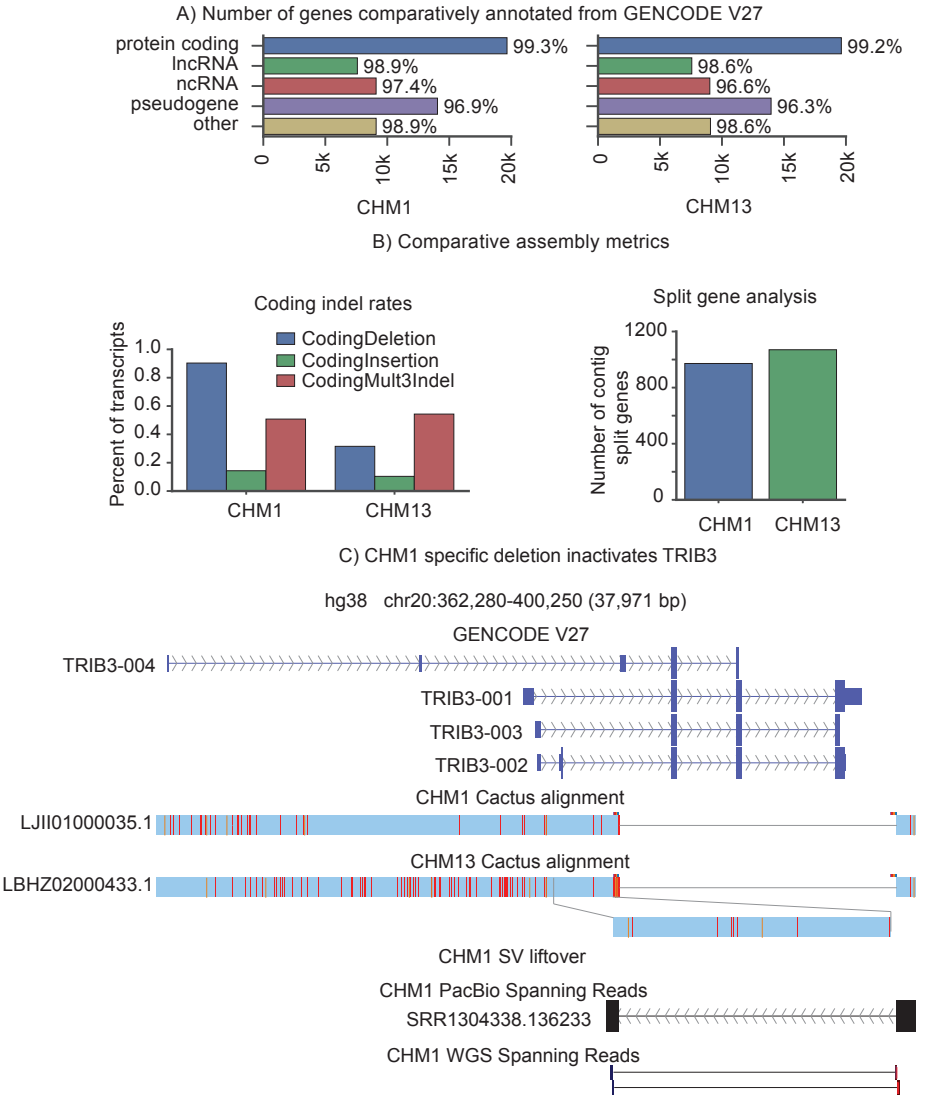


Figure 4.3: Pseudo-diploid human annotation

A) The number and fraction of genes comparatively annotated from GENCODE V27 in each assembly. GENCODE biotypes are simplified into protein coding, lncRNA, ncRNA, pseudogene and other. Other includes processed transcripts, nonsense-mediated decay, and immune-related genes. B) Comparative assembly metrics diagnose assembly quality. Frame shifting insertions, deletions and multiple of 3 indels that do not shift frame are reported for each assembly (left). Consistent with the great ape genomes, there is a systematic over-representation of coding deletions, despite these assemblies coming from haploid cell lines. Split gene analysis reports how often paralog-resolved transcript projections end up on different contigs, which can measure assembly gene-level contiguity. C) UCSC Assembly Hub[101] showing TRIB3 deletion in CHM1. Analysis of genes found in one genome and not the other led to the discovery of a novel structural variant specific to CHM1, which disables the gene TRIB3. Spanning reads were found in both PacBio and Illumina whole genome sequencing that validate the deletion.

A) Jaccard similarity in Rat annotation sets

annotation pair	Ensembl V90/RefSeq	CAT/ Ensembl V90	CAT/RefSeq	Ensembl V90/MAKER	CAT/ MAKER	MAKER/ RefSeq
exon Jaccard similarity	0.658	0.649	0.648	0.514	0.484	0.464
intron Jaccard similarity	0.749	0.74	0.841	0.364	0.334	0.337

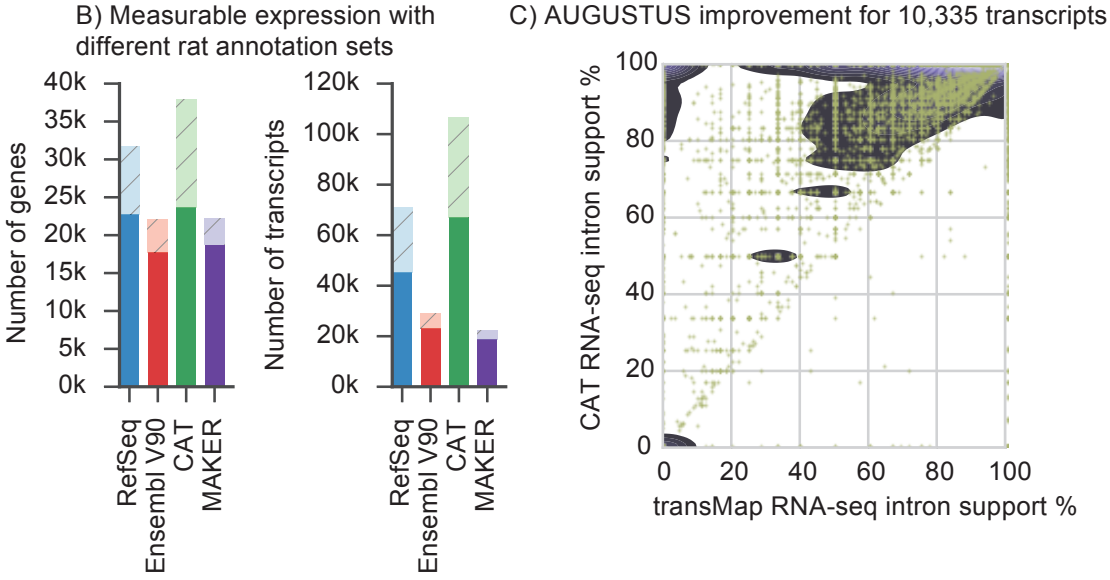
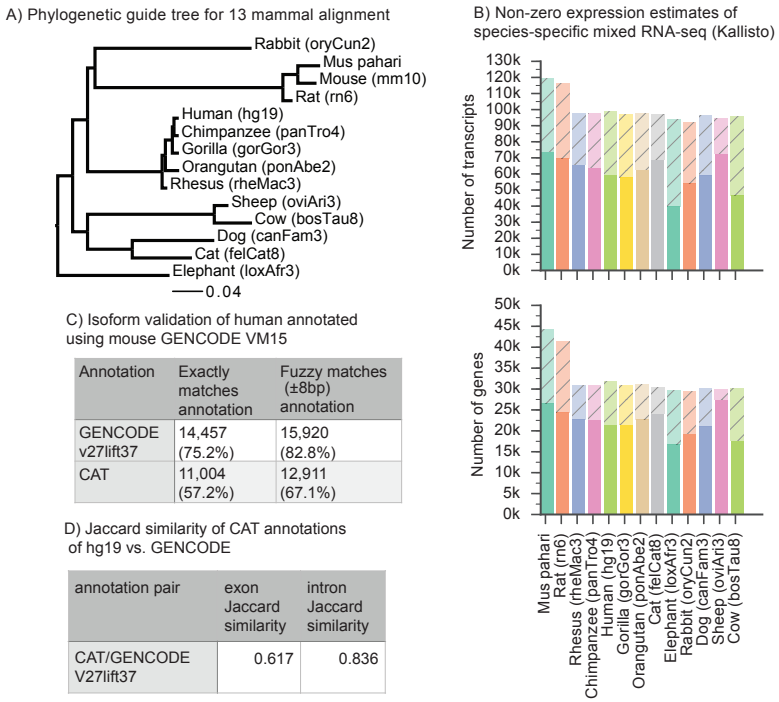


Figure 4.4: Validation of CAT annotation using rat

A) Jaccard similarity of CDS introns and exons between rat annotation sets shows high similarity between CAT and existing Ensembl and RefSeq annotations. B) Each transcript set was used to construct a Kallisto [156] index, and then all of the input RNA-seq for annotation were quantified. Solid bars are genes or transcripts with non-zero expression (TPM >0.1) estimates, while light hatched bars are the remainder of the annotation set. CAT provides an annotation set with slightly more detectable genes than other annotation methods, and far more detectable isoforms. C) AugustusTMR provides a mechanism to clean up transcript projections and shift splice sites, fixing alignment errors as well as real evolutionary changes. Most of the 10,335 AugustusTMR transcripts chosen in consensus finding show an improvement in RNA-seq support, which is one of the features used in consensus finding.



C) Isoform validation of human annotated using mouse GENCODE VM15

Annotation	Exactly matches annotation	Fuzzy matches (±8bp) annotation
GENCODE v27lift37	14,457 (75.2%)	15,920 (82.8%)
CAT	11,004 (57.2%)	12,911 (67.1%)

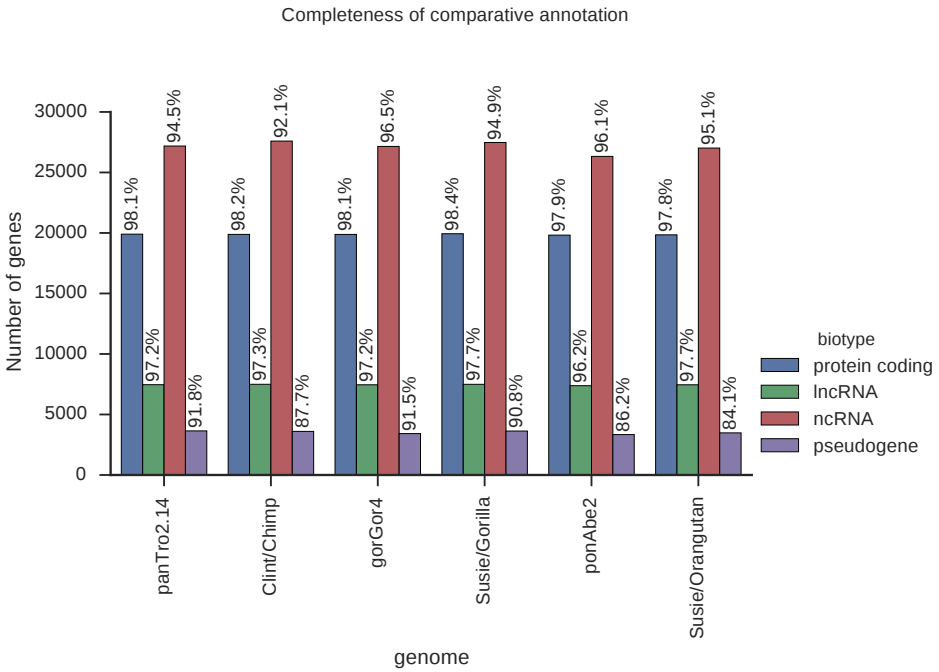
D) Jaccard similarity of CAT annotations of hg19 vs. GENCODE

annotation pair	exon Jaccard similarity	intron Jaccard similarity
CAT/GENCODE V27lift37	0.617	0.836

Figure 4.5: 13-way annotation

A) The phylogenetic guide tree for the 13-way mammal alignment. See the methods section for the exact Newick format tree. B) The gene annotation sets for each of the 13 mammalian genomes were quantified against the mixed input RNA-seq sets obtained from SRA. Genes or transcripts with TPM>0.1 are solid colors, while genes or transcripts with no measurable expression are shaded. An average of 2.8 isoforms per gene per genome had quantifiable expression, suggesting that CAT can infer isoform information across long branch lengths. C) ICE isoforms were compared in an exact and fuzzy matching scheme to both the GENCODE annotation of hg19 as well as the CAT annotation of hg19 derived from mouse. Even at large phylogenetic distance, CAT still correctly annotates 67.1% of isoforms identified by ICE. D) Jaccard similarity of CDS introns and exons in human annotation sets show that CAT identifies a larger portion of introns correctly than exons, with values similar to the mouse rat comparison.

A)


Completeness of comparative annotation
Number of genes
30000
25000
20000
15000
10000
5000
0
98.1%
97.2%
94.5%
91.8%
98.2%
97.3%
92.1%
87.7%
98.1%
97.2%
96.5%
91.5%
98.4%
97.7%
94.9%
90.8%
97.9%
96.2%
96.1%
86.2%
97.8%
97.7%
95.1%
84.1%
panTro2.14
Clint/Chimp
gorGor4
Susie/Gorilla
ponAbe2
Susie/Orangutan
genome
biotype
protein coding
lncRNA
ncRNA
pseudogene

B)

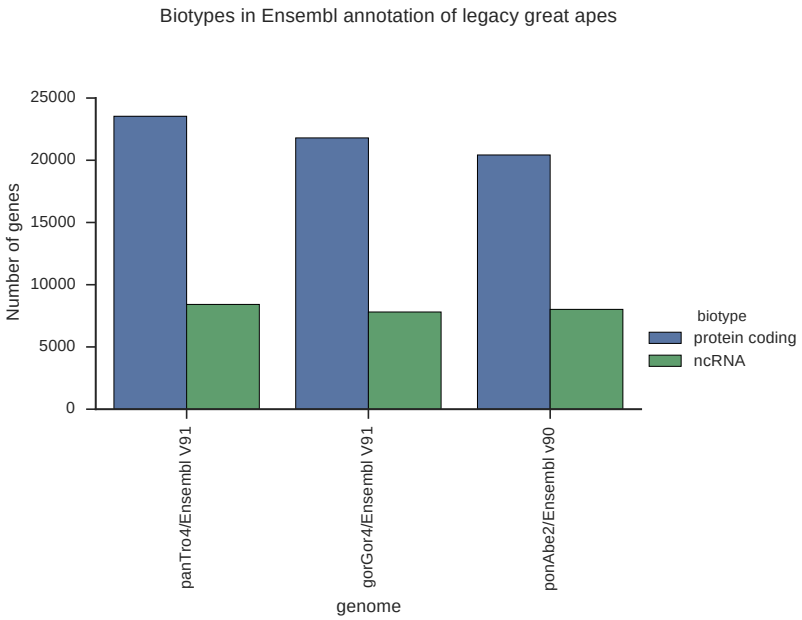


Figure 4.5: Primate completeness and biotypes

A) Percent of genes in simplified biotypes identified in both generations of great ape assemblies. The numbers above the bars are the percent of GENCODE V27 genes identified broken down by simplified biotypes. The number of genes identified in the PacBio assemblies increased slightly for all of the great apes. B) The biotypes of the Ensembl annotations for the older great apes. Compared to the 19,836 protein coding genes in GENCODE V27, these annotation sets have 23,534, 21,795 and 20,424 protein coding genes for chimpanzee, gorilla and orangutan respectively, suggesting misclassified non-coding loci. We found 940 loci in chimp, 1,728 loci in gorilla and 1,270 loci in orangutan which are labeled as protein coding in Ensembl but are labeled other biotypes in the CAT annotation. Not only does CAT make tracking orthology relationships easier, but it also provides much higher correlation with real data, greatly improving cross-species RNA-seq analysis.

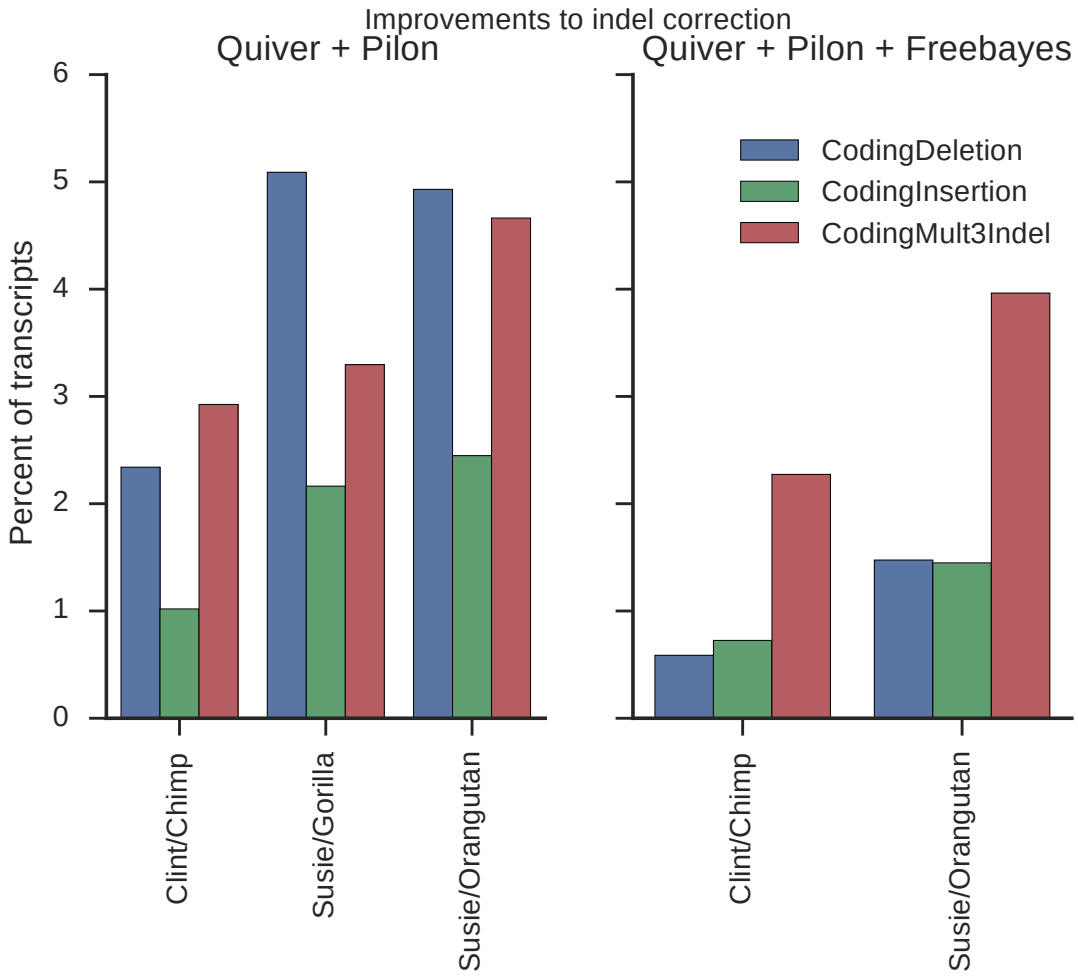


Figure 4.6: Primate Coding Indels

The rate of transcripts with coding indels seen in the consensus gene sets for the SMRT primate assemblies are shown with Quiver and Pilon correction (left), and subsequent Freebayes [83] based correction (right). Freebayes correction was not performed on GSMRT5 (gorilla). Coding indels are measured by pairwise translated BLAT alignments of a transcript to its ortholog in human. SMRT assemblies show a systematic over-representation of coding insertions. After Freebayes correction, the rate of insertions to deletions is roughly equal and lower than the rate of multiple of 3 indels, which is the expected result due to purifying selection.

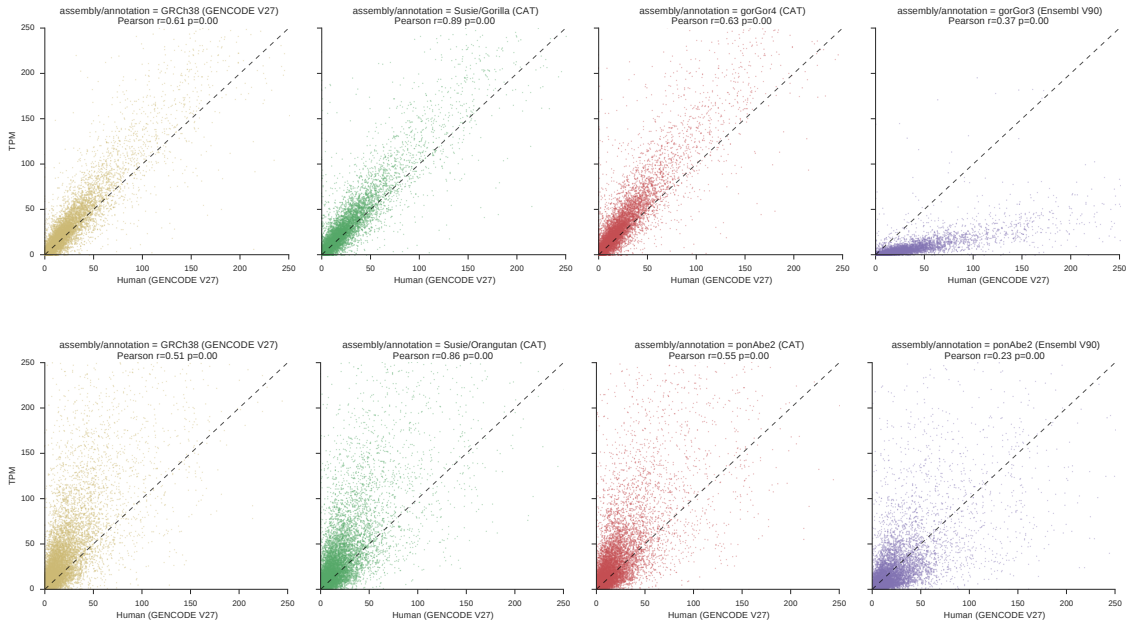
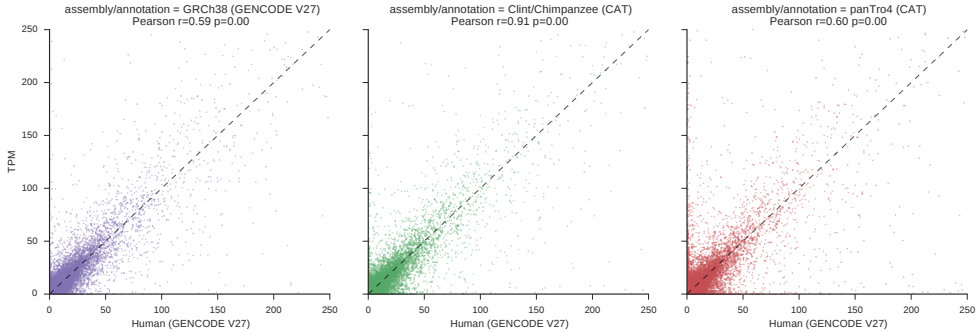


Figure 4.6: Cross-species RNA-seq expression estimates

Gorilla and orangutan iPSC RNA-seq are compared to human iPSC RNA-seq using a variety of annotation and assembly combinations. All comparisons were performed with Kallisto. Cross-species comparison was used by tracking gene common names, and only protein coding genes were considered. Because the pre-release of Ensembl V91 provided to us lacked common names, we used V90. Ensembl V90 annotation is on gorGor3, so that genome was used. The x-axis in all plots are the TPM of human iPSC data mapped to GENCODE V27. The y-axis in all plots is the TPM of species-specific iPSC RNA-seq mapped to the assembly/annotation pair in the title. In all cases, using the newest version of the assemblies with CAT provides the highest correlation.



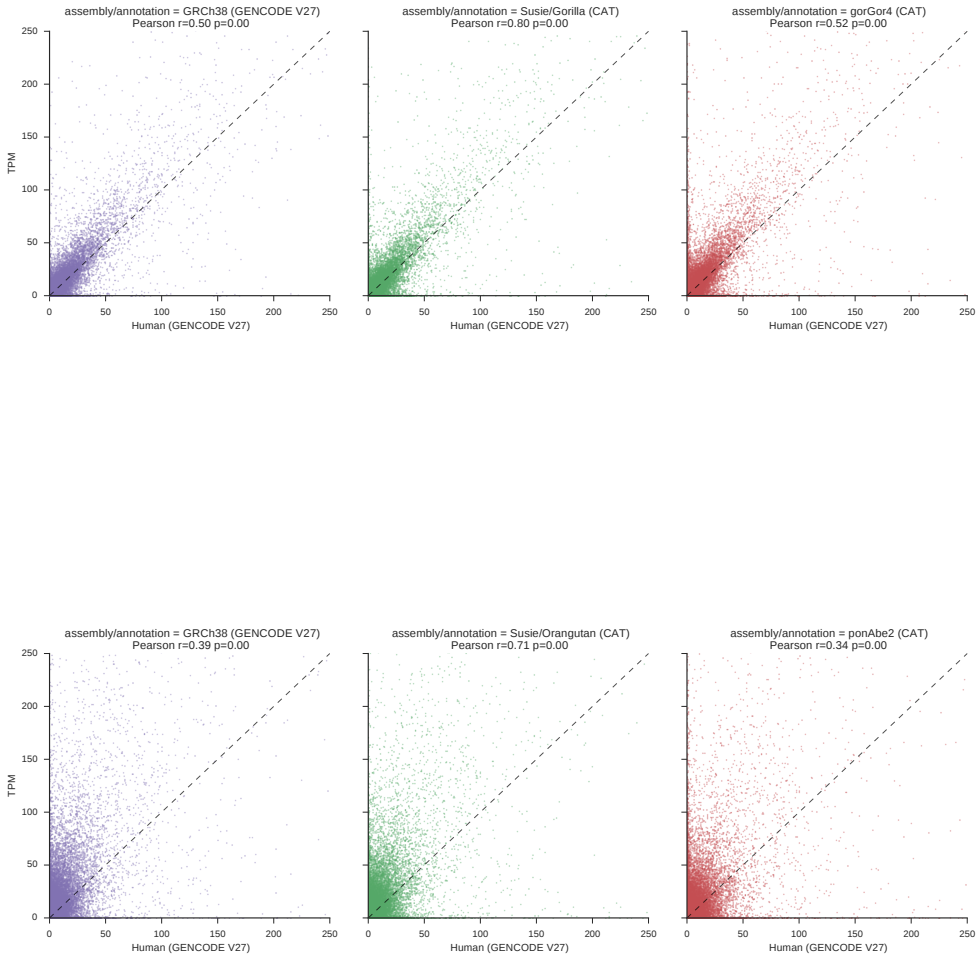


Figure 4.6: Cross-species RNA-seq isoform expression estimates

The same analysis as in Figure 4.2D and 4.6 was performed on the isoform level. For this analysis Ensembl was not included because we lacked a mapping of isoform IDs between the Ensembl annotation set and GENCODE. Only protein coding isoforms were considered. The highest correlation is seem in CAT annotation of SMRT genomes, although correlation falls off considerably as phylogenetic distance increases from chimpanzee to orangutan.

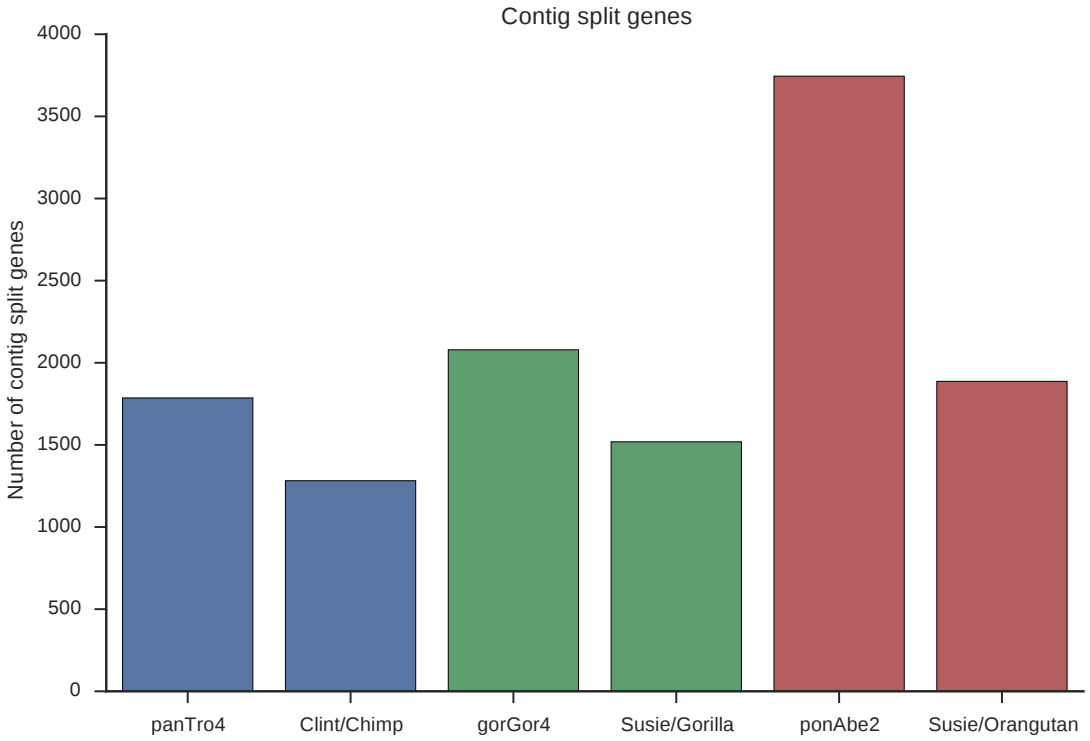


Figure 4.7: Primate Split Genes

Split gene analysis looks for transcript projections after paralog resolution that map to multiple contigs. This provides a metric for assembly contiguity. Despite the PacBio assemblies not being in chromosome sized pieces, fewer split genes are detected, suggesting that most contig breaks are not in genic regions.

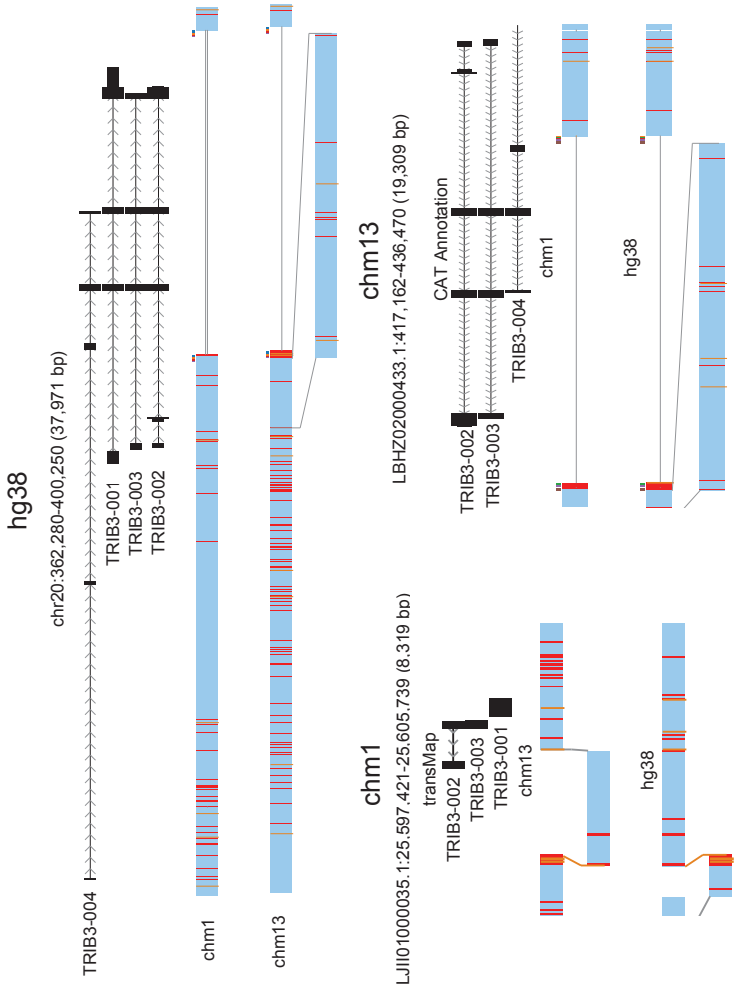


Figure 4.8: TRIB3 example

The CHM1 structural variant in figure 4 is shown here from all perspectives. In CHM1, the short transMaps for the few remaining exons are filtered out and do not end up in the annotation set. In contrast, CAT annotated 3/4 of the isoforms. This figure shows the power of the UCSC assembly hub for evaluating structural variants by being able to view the alignment from any species present.

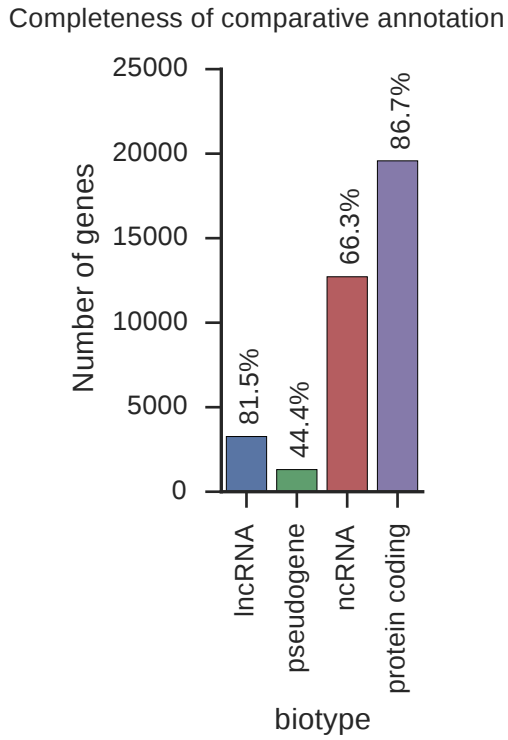


Figure 4.9: Rat completeness

The number of genes comparatively identified in the rn6 assembly from mm10 GENCODE VM11, broken down by simplified biotypes. The percentages on top are the percent of the total genes in each simplified biotype present in VM11. While a large portion of protein coding genes are identified, much fewer lncRNAs and other non-coding biotypes are identified.

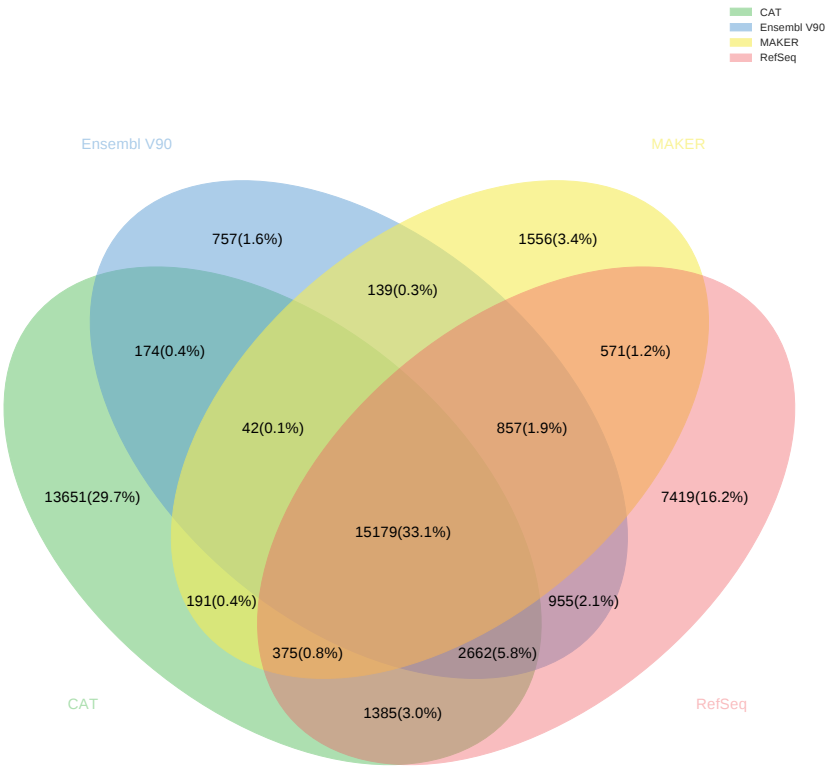
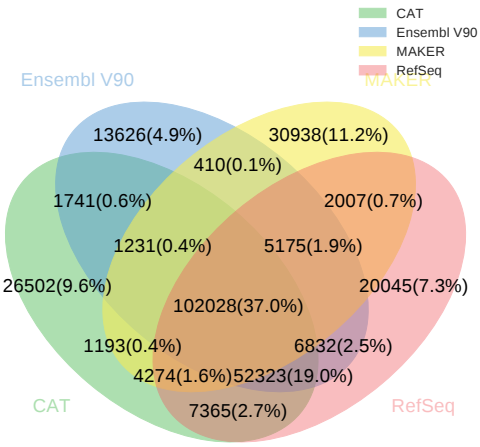
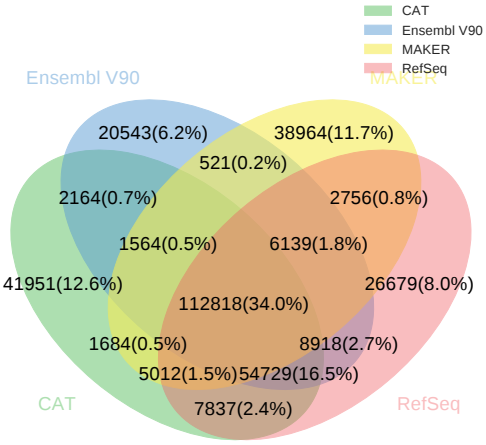


Figure 4.10: Rat Locus Venn Diagram

Gene loci were compared between CAT, Ensembl V90, MAKER and RefSeq on rat rn6. Loci were clustered using the Kent tool clusterGenes, which requires exonic overlap on the same strand. Only 15,179 loci are shared between all sets.



A)



B)

Figure 4.10: Rat Exon/Intron Support Venn Diagram

CDS Intron (left) and CDS exon (right) interval exact matches were compared between CAT, Ensembl V90, MAKER and RefSeq on rat rn6. MAKER had the highest proportion of unsupported exons and introns, followed by CAT. Only 37.0% and 34.0% of introns and exons respectively are present in all four annotation sets.

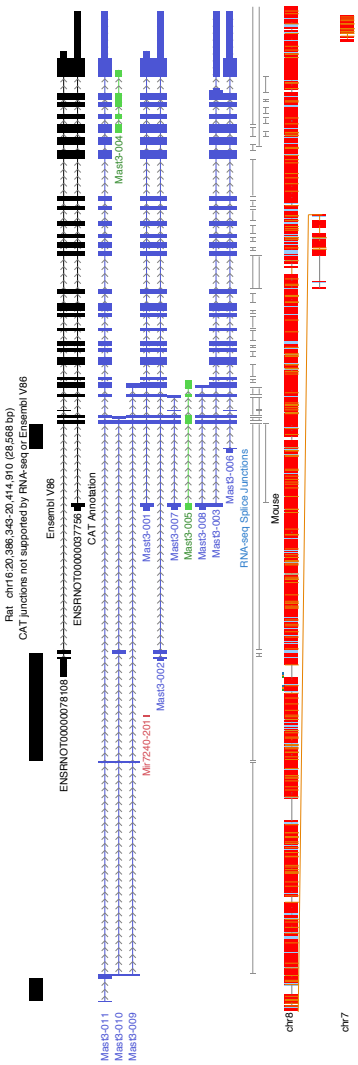


Figure 4.11: Unsupported junctions example

The rat gene Mast3 has two annotated isoforms in Ensembl supported by RNA-seq. CAT annotation added 9 new isoforms, two of which had unsupported junctions. These new annotations reveal an upstream transcription initiation site supported by RNA-seq.

Genome	Number of BUSCO genes missing
Mus pahari	38
Rat (rn6)	99
Rhesus (rheMac3)	138
Chimpanzee (panTro4)	90
Human (hg19)	26
Gorilla (gorGor3)	184
Orangutan (ponAbe2)	133
Cat (felCat8)	95
Elephant (loxAfr3)	114
Rabbit (oryCun2)	147
Dog (canFam3)	90
Sheep (oviAri3)	162
Cow (bosTau8)	93

Table 4.1: BUSCO genes missing in 13 mammal annotation

BUSCO was used to quantify the number of core key genes missing in the CAT annotation of 13 mammalian genomes. For this analysis, the mammalian odb_9 set of 4,104 genes was used. BUSCO genes represent core housekeeping genes present at single copy across long evolutionary distance. On average, 108 BUSCO genes (2.63%) are missing in each genome. Only three BUSCO genes (EOG090A0GHJ, EOG090A05ND, and EOG090A04MN) were missing in all 13 genomes.

Table 4.2: SRA RNA-seq accessions

Species	SRA Accessions	Tissues
Rat	SRR1041777, SRR1768421, SRR1768443, SRR1768444, SRR299123, SRR636875, SRR636876, SRR636877, SRR636925, SRR636926, SRR636927, SRR636970, SRR636971, SRR636972	Mixed, testis, liver, kidney, brain
Orangutan	SRR306792, SRR2176206, SRR2176207	Brain, testis
Gorilla	SRR832925, SRR3053573, SRR306809, SRR306803, SRR306804, SRR306801, SRR306807, SRR306810, SRR306805, SRR306806, SRR306802, SRR306800, SRR306808	Brain, 20 tissue pool
Chimp	SRR2040584, SRR2040585, SRR2040586, SRR2040587, SRR2040588, SRR2040589, SRR2040590, SRR2040591, SRR3711187, SRR3711188, SRR873622, SRR873623, SRR873624, SRR873625	brain, heart, liver, testis, 8 week old iPSC derived neurons, undifferentiated iPSC
Rhesus	SRR306784, SRR306786, SRR306785, SRR2040593, SRR306783, SRR306787, SRR306780, SRR306778, SRR306790, SRR2040595, SRR2040594, SRR2040592, SRR306788, SRR306782, SRR306789, SRR306777, SRR306779, SRR306781	Kidney, liver, heart, brain, testis
Human	ERR579132, ERR579133, ERR579134, ERR579135, ERR579136, ERR579137, ERR579138, ERR579139, ERR579140, ERR579141, ERR579142, ERR579143, ERR579144, ERR579145, ERR579146, ERR579147, ERR579148, ERR579149, ERR579150, ERR579151, ERR579152, ERR579153, ERR579154, ERR579155	Ovary, tonsil, fallopian tube, placenta, endometrium, rectum, skeletal muscle, liver, fat, colon, smooth muscle, lung

Sheep	SRR1653601, SRR1561187, SRR1561150, SRR1265856, SRR1536790, SRR1561171, SRR1265854, SRR1561367, SRR1561365, SRR1653570, SRR1653598, SRR1653597, SRR1266019, SRR1265849, SRR1653600, SRR1656805, SRR1561366, SRR1653594, SRR1561196, SRR1265855, SRR1653596, SRR1536788, SRR1266022, SRR1561156, SRR1266018, SRR1561195, SRR1536770, SRR1266020	Liver, brain, blood
Cow	SRR2960011, SRR2960020, SRR2960008, SRR2960010, SRR2960012, SRR2960016, SRR2960006, SRR2960015, SRR2960025, SRR2960003, SRR2960022, SRR2960029, SRR2960030, SRR2960017, SRR2960032, SRR2960005, SRR2960027, SRR2960007, SRR2960036, SRR2960026, SRR2960035, SRR2960004, SRR2960002, SRR2960034, SRR2960013, SRR2960001, SRR2960021, SRR2960019, SRR2960009, SRR2960024, SRR2960014, SRR2960031, SRR2960033, SRR2960023, SRR2960028, SRR2960018	Liver, udder
Elephant	SRR1041765, SRR975189, SRR975188, SRR3222430	Fibroblast
Rabbit	SRR636919, SRR636872, SRR636964, SRR636871, SRR636920, SRR636965	Liver, kidney, brain
Cat	SRR3200450, SRR3200448, SRR3200449, SRR3200453	Fetus, lung, liver

Publicly available RNA-seq obtained via SRA for annotations performed in this paper.

Chapter 5

Usages of CAT

This chapter provides a brief overview of the papers that CAT has been applied to. For each of these papers, at a minimum, I ran CAT on the genomes in question and provided analysis of the results. In many of these projects I informed on assembly metrics and helped guide other groups on how to improve their assemblies, leading to improved gene annotations.

5.1 Overview

CAT has been used in a wide variety of projects at this point. After the initial work in the mouse genomes project, CAT has been applied to the following projects:

1. Repeat associated mechanisms of genome evolution and function revealed by the *Mus caroli* and *Mus pahari* genomes [130].
2. Long read assembly of the gorilla genome [40].
3. Nanopore sequencing and assembly of a human genome with ultra-long reads [128].
4. Revisiting the mouse reference genome: single molecule sequencing of C57BL/6J 'Eve' (publication pending).
5. Genome Assembly and Renal Gene Expression Before and After Hibernation in the American Black Bear (submitted).
6. NHGRI Reference Genomes Improvement Project.

7. High-resolution comparative analysis of great ape genomes (publication pending).

8. Structurally conserved primate lncRNAs are transiently expressed during human cortical differentiation and influence cell type specific genes (submitted).

9. A 3-way hybrid approach to generate a new high quality chimpanzee reference genome (Pan_tro_3.0) (accepted).

Here, I will provide a brief overview of my contribution to these projects.

5.2 Primate Genomes

I became involved in the gorilla long read assembly project [40] in late 2015. This project produced one of the first mammalian genomes sequenced with PacBio and assembled with Falcon. I ran comparativeAnnotator on this genome and was able to report an extremely high rate of coding indels - 63% of protein coding transcripts had some sort of frame-shifting indel. This led the Eichler lab to go and perform Illumina based error correction on the genomes, and the final published genome had a 26% frame-shifting indel rate. This is still far higher than the real biological rate.

5.3 *Mus pahari* and *Mus caroli*

Mus pahari and *Mus caroli* are two wild-derived inbred laboratory strains of mice that were included in the cactus alignments and annotation of the Mouse Genomes Project. I ran comparativeAnnotator on these genomes, and the annotation sets for these genomes are now part of Ensembl.

5.4 lncRNAs in primates

I applied CAT to a alignment of GRCh37, panTro4, gorGor4, ponAbe2, rheMac8 and other outgroup genomes. I then took the FANTOM lv3 lncRNA annotation set in human and lifted it to all of these genomes. I then further made use of the homGeneMapping tool to project RNA-seq information at various timepoints in cortical

organoid differentiation to all of these genomes. This analysis provided a whole-genome view of lncRNA conservation across great apes.

5.5 Eve

I applied CAT to a *de-novo* assembly of a PacBio assembly of Eve, a lab mouse strain derived from very early frozen embryos of the C57BL/6J strain, which was the strain used for the reference mouse assembly (mm10). This project was recently presented as as a poster at ASHG.

5.6 NHGRI Reference Genomes Project

The NHGRI Reference Genomes Project is a project to produce high quality PacBio assemblies of representatives of various human populations. I used CAT to annotate 5 such genomes and inform on the indel error rates, leading to the use of improved indel correction. This project is ongoing.

5.7 Nanopore seqencing of NA12878

I applied CAT in this project to annotate the Oxford Nanopore based assembly of NA12878, a well studied human individual. Indel analysis showed that this genome assembly has a similar level of errors to uncorrected PacBio assemblies, and has very high gene locus contiguity.

Chapter 6

NOTCH2NL

6.1 Introduction

In my time at UCSC, I also worked on a very different comparative genomics project. The NOTCH2NL project is an intensive effort to understand a complicated, segmentally duplicated region of human chromosome 1 implicated in human neurodevelopmental phenotypes.

I was the lead author on this project. I developed most of the genomics experiments represented in this paper, including designing the microarray based enrichment approach to analyzing normal and patient genomes. Sections in the methods that I developed are the 10x Library Enrichment, NOTCH2NL Simons Samples Coverage Analysis, Hominid Copy Number Analaysis, Gordian Assembler (in conjunction with Alex Bishara), Establishment of Paratypes in Population (in conjunction with Colleen Bosworth), Enrichment and Sequencing of Full-Length cDNA (in conjunction with Gary Mantalas and Ryan Lorig-Roach), CRISPR Mutation of NOTCH2NL in the H9 ES Line (in conjunction with Meghan Mooring), Estimate of NOTCH2 and NOTCH2NL Expression in human fetal brain scRNA-Seq (in conjunction with Aparna Bhadun), Paraype Estimation of NOTCH2NL in Human Population, and NOTCH2NL Expression in Week 5 Neurospheres (in conjunction with Frank Jacobs and Andrew Field).

I wrote a significant portion of the text and helped in editing the entire paper. I would like to thank David Haussler for his dedication to this project and to making the impossible possible.

This project is in many senses very similar to CAT. It is a comparative genomics project that relies on accurate assembly and annotation of genes to make biological sense of data. It is different in that instead of being a genome-wide scan, it is focused on a specific location and gene family.

Below is the full paper as submitted to *Cell*, and published to BioRxiv as of November 18th, 2017.

6.2 NOTCH2NL paper

Human-specific *NOTCH*-like genes in a region linked to neurodevelopmental disorders affect cortical neurogenesis

Authors:
Ian T Fiddes[1,12], Gerrald A Lodewijk[2,12], Meghan Mooring[1], Colleen M Bosworth[1], Adam D Ewing[1#], Gary L Mantalas[1,3], Adam M Novak[1], Anouk van den Bout[2], Alex Bishara[4], Jimi L Rosenkrantz[1,5], Ryan Lorig-Roach[1], Andrew R Field[1,3], Maximilian Haeussler[1], Lotte Russo[2], Aparna Bhaduri[6], Tomasz J. Nowakowski[6], Alex A. Pollen[6], Max L. Dougherty[7], Xander Nuttle[8], Marie-Claude Addor[9], Simon Zwolinski[10], Sol Katzman[1], Arnold Kriegstein[6], Evan E. Eichler[7,11], Sofie R Salama[1,5,13], Frank MJ Jacobs[1,2,13.14]*, David Haussler[1,5,13,14]*

Affiliations:
[1] UC Santa Cruz Genomics Institute, Santa Cruz, California, United States of America,
[2] University of Amsterdam, Swammerdam Institute for Life Sciences, Amsterdam, The Netherlands
[3] Molecular, Cell and Developmental Biology, of California Santa Cruz, Santa Cruz, California, United States of America
[4] Department of Computer Science and Department of Medicine, Division of Hematology, Stanford University, California, USA
[5]Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, California, United States of America
[6]The Eli and Edythe Broad Center for Regeneration Medicine and Stem Cell Research at UCSF, San Francisco, USA
[7]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA
[8]Center for Genomic Medicine, Massachusetts General Hospital, Harvard Medical School, and Broad Institute of MIT and Harvard, Boston, Massachusetts, United States of America
[9]Service de génétique médicale, Lausanne, Switzerland
[10] Department of Cytogenetics, Northern Genetics Service, Institute of Genetic Medicine, Newcastle upon Tyne, United Kingdom
[11]Howard Hughes Medical Institute, University of Washington, Seattle, WA, United States of America
[12] These authors contributed equally to this work
[13] Senior authors
[14] Lead contact
* corresponding authors: haussler@ucsc.edu, F.M.J.Jacobs@uva.nl
[#]Present address: Translational Research Institute, Princess Alexandra Hospital, Brisbane, Australia

Summary:

Genetic changes causing dramatic brain size expansion in human evolution have remained elusive. Notch signaling is essential for radial glia stem cell proliferation and a determinant of neuronal number in the mammalian cortex. We find three paralogs of human-specific NOTCH2NL are highly expressed in radial glia cells. Functional analysis reveals different alleles of NOTCH2NL have varying potencies to enhance Notch signaling by interacting directly with NOTCH receptors. Consistent with a role in Notch signaling, NOTCH2NL ectopic expression delays differentiation of neuronal progenitors, while deletion accelerates differentiation. NOTCH2NL genes provide the breakpoints in typical cases of 1q21.1 distal deletion/duplication syndrome, where duplications are associated with macrocephaly and autism, and deletions with microcephaly and schizophrenia. Thus, the emergence of hominin-specific NOTCH2NL genes may have contributed to the rapid evolution of the larger hominin neocortex accompanied by loss of genomic stability at the 1q21.1 locus and a resulting recurrent neurodevelopmental disorder.

Introduction

The human brain is characterized by a large neocortex that forms the substrate for the development of human-specific higher cognitive functions (Lui et al., 2011; Molnar et al., 2006; Rakic, 2009), but evolutionary changes to our genome that underlie the increase in size and complexity of the human neocortex are poorly understood (Varki et al., 2008). Structural genomic variants account for 80% of human-specific base pairs and are therefore an important class of genomic regions to consider (Cheng et al., 2005). Of

particular interest are loci where segmental duplications have created entirely new human-specific gene paralogs that are associated with cortical development, such as *SRGAP2C* (Dennis et al., 2012), (Charrier et al., 2012), *ARHGAP11B* (Florio et al., 2015) and *TBC1D3* (Ju et al., 2016). Interestingly, human-specific duplicated genes are often located within segmental duplications that mediate recurrent rearrangements associated with neurodevelopmental disorders (Stankiewicz and Lupski, 2010) (Florio et al., 2015), (Nuttle et al., 2016), (Popesco et al., 2006), (Dumas et al., 2012), (Dougherty et al., 2017). One region susceptible to these rearrangements lies on human chromosome band 1q21, which was involved in a large pericentric inversion involving considerable gene loss and duplication during human evolution (Szamalek et al., 2006), contains a disproportionate number of human-specific genes (O'Bleness et al., 2012), and also contains the 1q21.1 distal deletion/duplication syndrome interval (Mefford et al., 2008), (Brunetti-Pierri et al., 2008). *De novo* deletion of one copy of this 1q21.1 locus frequently leads to an abnormal reduction in brain size (microcephaly) and reciprocal duplication often results in an abnormal increase in brain size (macrocephaly), among other symptoms.

The 1q21.1 locus was not correctly assembled in the human reference genome until the most recent version, GRCh38, hampering early research of 1q21.1 syndromes. This is due to several tracts of almost identical paralogous DNA. The most recent work (Dougherty et al., 2017) hypothesized that the typical breakpoints for the distal 1q21.1 deletion/duplication syndrome lie within two of these tracts, at ~chr1:146,200,000 and chr1:148,600,000 in GRCh38, respectively, in regions of ~250 kb at 99.7% identity. We confirm this, and show that these regions contain little-studied, paralogous genes we

call *NOTCH2NLA* and *NOTCH2NLB* that affect Notch signaling during human neurodevelopment. There are also atypical breakpoints for the distal 1q21.1 syndrome (Dougherty et al., 2017) that reach further almost to a third *NOTCH2NL* paralog we call *NOTCH2NLC.* We show both the typical and atypical breakpoints lead to change in NOTCH2NL copy number. This suggests that *NOTCH2NL* paralogs could contribute to the brain size and other cognitive symptoms of 1q21.1 distal deletion/duplication syndrome since Notch signaling is central to brain development, determining the timing and duration of neuronal progenitor proliferation and neuronal differentiation (Louvi and Artavanis-Tsakonas, 2006), (Kageyama et al., 2009), (Hansen et al., 2010).

We also find that *NOTCH2NL* genes are unique to the human, Neanderthal and Denisovan species among all species that have been sequenced (i.e., unique to hominin species), and originated just prior to or during the period where fossil evidence demonstrates the most rapid growth in the size of the hominin neocortex (Holloway et al., 2004). *NOTCH2NL*'s origin appears to be in an ectopic gene conversion event 3-4 million years ago in which part of *NOTCH2* "overwrote" an earlier *NOTCH2*-like pseudogene creating a new functional, short, secreted form of *NOTCH2*.

NOTCH2NL is expressed in the developing fetal brain at highest levels in radial glia neural stem cells, including outer radial glia cells, a cell type hypothesized to generate the majority of primate cortical neurons and to contribute more to cortical expansion in hominin than in other apes (Lewitus et al., 2013), (Smart et al., 2002). Functional analysis of *NOTCH2NL* by ectopic expression in mouse embryonic stem cell (mESC)-derived cortical organoids, as well as analysis of CRISPR-generated *NOTCH2NL* deletion mutant human embryonic stem cell (hESC)-derived cortical organoids,

suggests that *NOTCH2NL* expression delays differentiation of neuronal progenitor cells including dorsal radial glial cells. This is consistent with cerebral neoteny observed in hominin brain development where, perhaps for obstetrical reasons, the human and Neanderthal brains are slower to mature during fetal development and expand by a factor of 3.3 after birth, compared to a factor of 2.5 in chimpanzees (DeSilva and Lesnik, 2006; Ponce de Leon et al., 2008).

Finally, we demonstrate that NOTCH2NL physically interacts with NOTCH receptors and can activate NOTCH signaling in reporter assays in the presence and absence of NOTCH ligands. From functional and genomic evolutionary evidence, we propose that with the creation of the modern forms of *NOTCH2NL* genes in the last few million years, hominins gained a new, secreted NOTCH-like protein that can act to amplify Notch signaling and promote increased cortical neurogenesis by delaying differentiation of neuronal progenitor cells. Thus, the emergence of *NOTCH2NL* genes in hominins may have contributed to the increase in size and complexity of the human neocortex at the expense of susceptibility to 1q21.1 distal duplication/deletion syndrome.

Results

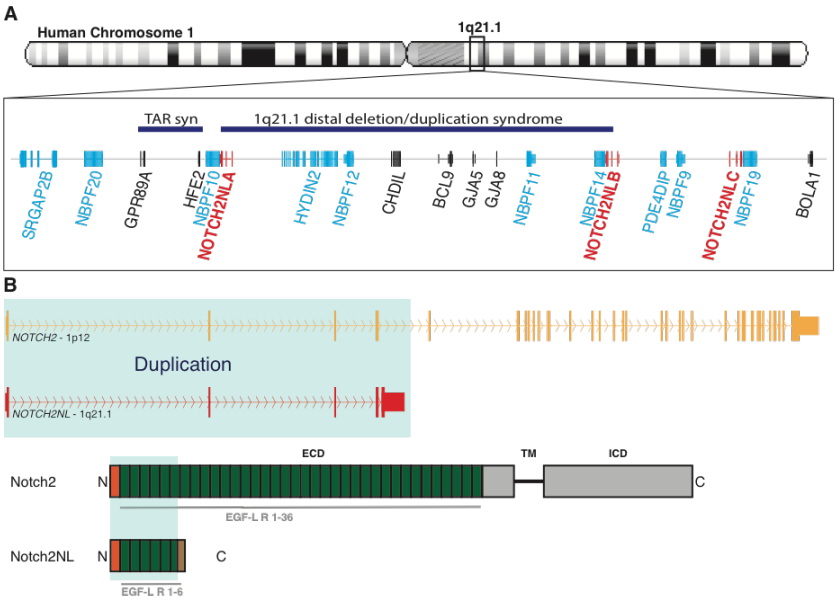
NOTCH2NL is a novel NOTCH-like gene, uniquely expressed in humans

The gene annotated as *NOTCH2NL* (Duan et al., 2004) on human genome assembly GRCh37 resides on the q arm of human chromosome 1 in the 1q21.1 locus, which long remained one of the most difficult regions in our genome to assemble due to its highly repetitive nature (Szamalek et al., 2006), (Dennis et al., 2012), (Doggett et al., 2006). The 1q21.1 locus has undergone a number of human lineage-specific rearrangements,

including a pericentric inversion (Szamalek et al., 2006) and the creation of new human-specific gene paralogs, including *HYDIN2* (Dougherty et al., 2017) and *SRGAP2B* (Dennis et al., 2012). Resequencing of the pericentric region of chromosome 1 in a haploid human cell line finally resolved previously unmapped regions and led to a revised assembly of the 1q21.1 locus, which is incorporated in the human genome assembly GRCh38 (Steinberg et al., 2014). This improved assembly reveals the presence of four paralogous *NOTCH2NL*-like genes (**Figure 1A**): *NOTCH2NLA*, *NOTCH2NLB* and *NOTCH2NLC* reside in the 1q21.1 locus, and a fourth quite different paralog, *NOTCH2NLR* (NOTCH2NL-Related) is located near *NOTCH2* on the p-arm of chromosome 1. The greater than 100 kb genomic regions spanning each *NOTCH2NL* gene show greater than 99.1% identity to *NOTCH2* (**Figure S1A**), suggesting that *NOTCH2NL* paralogs were created within the last few million years, on a time scale similar to that of *SRGAP2* and *HYDIN2* (Dennis et al., 2012), (Dougherty et al., 2017).

NOTCH2NL results from a partial duplication of the *NOTCH2* gene. The duplicated segment includes the upstream region of *NOTCH2*, the *NOTCH2* promoter and EGF domains from *NOTCH2* exons 1 through 4, while excluding the membrane spanning and cytoplasmic domains of *NOTCH2*. Because it lacks a transmembrane domain, the NOTCH2NL protein can be secreted by the cells that express it (Duan et al., 2004). *NOTCH2NL* genes contain a fifth exon derived from *NOTCH2* intronic sequence that provides NOTCH2NL with 20 unique amino acids (**Figure 1B, Figure S1I**). In paralogs *NOTCH2NLA, NOTCH2NLB* and *NOTCH2NLC* the 5th exon has a 4 bp deletion compared to the corresponding sequence in *NOTCH2*. We established that this 4 bp

deletion is essential for NOTCH2NL protein expression (**Figure S1B-E**). *NOTCH2NLR* lacks this 4 bp deletion, contains many variants in its protein coding region relative to



C - *NOTCH2NL* Genotype in H9 hESCs

Loc.	Paratype	NOTCH2NL Protein Domains	Relative Expression
A/B	NOTCH2NL^{Sh}		16.1%
A/B	NOTCH2NL^{Sh,T197I}		17.0%
A/B	NOTCH2NL^{L}		43.6%
A/B	NOTCH2NL^{L,R113*}		3.5%
C	NOTCH2NL$^{Sh\text{-}2ntdel}$		9.9%*
C	NOTCH2NL$^{Sh\text{-}2ntdel}$		9.9%*

EGF-like domain; Secratory pathway signal sequence; * Premature stop codon; Conserved Thr fucosylation site; Truncated EGF-like domain; EGF-like domain with Ala to Thr mutation

D - *NOTCH2NL* in Human Population

Paratype	NOTCH2NL Protein Domains	Number of copies observed (n = 72)
NOTCH2NL^{L}		6
NOTCH2NL^{L,T197I}		5
NOTCH2NL^{Sh}		21
NOTCH2NL$^{Sh\text{-}2ntdel}$		24
NOTCH2NL$^{Sh\text{-}A154T}$		1
NOTCH2NL^{Sh,T197I}		12
NOTCH2NL^{L,R113*}		2

***Figure 1. NOTCH2NL* genes are associated with a neurodevelopmental disease locus and exhibit variable protein structures in the human population.** (A) Location of *NOTCH2NL* genes in chromosome 1q21.1 flanking the 1q21.1 distal deletion/duplication syndrome locus. This region contains many additional genes derived from human segmental duplication (shaded light blue). TAR syn is Thrombocytopenia Absent Radius syndrome. (B) Gene and protein features of NOTCH2 and NOTCH2NL. (C) Results of the *de novo* assembly of *NOTCH2NL* loci for H9 human embryonic stem cells and relative allele expression from week 5 cortical organoids as measured by full-length cDNA sequencing. * Not enough nucleotide differences present to distinguish between the two *NOTCH2NL*^{sh2ntdel}

alleles, so counts evenly split between the two. (D) Observed *NOTCH2NL* paratypes (n=72, counting paratypes transmitted from parent to child only once) obtained from linked-read sequencing and assembly of 15 individuals. See also Fig. S1 and Table S1 and S4.

NOTCH2 and *NOTCH2NL* paralogs including a very recent still segregating variant (**Figure S1I**), and analysis of human genome sequence data from the Simons diversity project (N=266) (Mallick et al., 2016) reveal a significant fraction of the population (14%) lack *NOTCH2NLR* (**Figure. S1G**). Together these results suggest that *NOTCH2NLR* is a non-functional pseudogene, although additional investigation would be needed to verify that its mRNA and/or protein product is not functional.

Despite the overall high sequence homology of the *NOTCH2NL*-gene loci as present on GRCh38, each *NOTCH2NL* gene paralog has distinguishing characteristics (**Figure 1D**): *GRCh38NOTCH2NLA* has an ATG→ ATA mutation in the *NOTCH2* initiator methionine codon and as a result, GRCh38NOTCH2NLA protein lacks the N-terminal 39 amino acids including the signal peptide of NOTCH2. We term this allele “short NOTCH2NL” and denote it *NOTCH2NL*^{Sh}. *GRCh38NOTCH2NLB* encodes a longer protein with the signal peptide, but has a Thr→Ile substitution in the 5th EGF repeat, and is referred to as “long NOTCH2NL with Thr→Ile substitution” (*NOTCH2NL*^{L,T197I}, coordinates relative to the NOTCH2 protein). This substitution occurs at a site for fucosylation of NOTCH2. *GRCh38NOTCH2NLC* contains a 2 bp deletion just downstream the original NOTCH2 start codon (*NOTCH2NL*$^{Sh\text{-}2ntdel}$), and like *NOTCH2NL*^{Sh}, NOTCH2NLC protein is initiated at a downstream ATG, resulting in a short NOTCH2NL protein that lacks the N-terminal signal peptide.

To better understand the spectrum of *NOTCH2NL* alleles in the human population we sought to fully resolve the *NOTCH2NL* haplotypes in several individuals. To overcome

the challenge of accurately assembling these highly homologous genes we took advantage of the linked reads of readset genome sequence technology whereby small DNA molecules compatible with short read DNA sequencers are labeled (barcoded) according to individual long genomic DNA fragments (>30 kb) from which they are derived. We developed an assembly-by-phasing approach called Gordian Assembler utilizing this barcoded data to resolve the five NOTCH2-related gene paralogs in 8 normal individuals (STAR Methods, **Figure S1H**, **Table S1**). This and related analysis revealed recent, likely ongoing, ectopic gene conversion occurring between *NOTCH2NLA*, *B* and *C* (**Figure S1F**) that is so extensive that *NOTCH2NLA* and *NOTCH2NLB* act essentially as a single gene with 4 alleles, and to some extent all three paralogs act as a single gene with 6 alleles. Our discovery of a large number of gene conversion alleles segregating in the human population is consistent with previous analysis of 1q21.1 (Nuttle et al., 2013). This, coupled with ~250 kb of DNA at 99.7% identity between the genomic regions containing *NOTCH2NLA* and *NOTCH2NLB*, makes it nearly impossible to assign individual alleles to a specific locus. Focusing on one assembly, that of the H9 hESC line, we identified six different alleles derived from *NOTCH2NLA*, *NOTCH2NLB* and *NOTCH2NLC* (**Figure 1C, Figure S1F**). These include three additional *NOTCH2NL* alleles beyond those present in the GRCh38 reference genome, including a short version with Thr→Ile (*NOTCH2NL*^{Sh,T197I}) and a long version without Thr→Ile (*NOTCH2NL*^{L}). We also observed an additional long version with a SNP (rs140871032) causing an in-frame stop codon at amino acid 113 (*NOTCH2NL*^{L,R113*}), also seen in 35 of 266 genomes in the Simons Diversity collection (Mallick et al., 2016). Finally, we observed a rare paratype *NOTCH2NL*^{Sh,A154T}

(rs76765512) with a non-synonymous change in EGF repeat 3 in one assembled genome and six Simons Diversity collection genomes. By RNA-seq analysis of cortical organoids derived from H9 hESCs (see below), we confirmed that all the *NOTCH2NL* alleles assembled in H9 are expressed. *NOTCH2NL*^{L,R113*} represents only about 3.5% of the transcripts, perhaps due to nonsense-mediated decay (**Figure 1C**). In the 17 genomes in which we have determined 72 *NOTCH2NL* alleles, we have found seven different *NOTCH2NL* sequence variants that affect the features of the resulting protein (**Figure 1D, Figure S1I**). We refer to these various *NOTCH2NL* alleles as paratypes since their physical location can vary among paralogous locations and therefore they do not conform to standard haplotypes. A typical *NOTCH2NL* genotype consists of 6 *NOTCH2NL* paratypes (**Figure 1C**), rather than two haplotypes as for most genes. Sequencing of these initial 72 *NOTCH2NL*-related paratypes is unlikely to have revealed all the structural variation in *NOTCH2NL* in the human population. We expect more variants will be found as additional genomes are sequenced and assembled in a manner allowing us to read these paratypes (**Figure 1D**, **Table S1, S4**).

Multiple rounds of gene duplication and gene conversion lead to functional *NOTCH2NL* genes only in hominins

To explore the evolutionary history of *NOTCH2NL* genes in more detail, we assessed the presence and structure of *NOTCH2NL* genes in other primate genomes. Based on alignment of genomic DNA-reads matching or nearly matching the parental *NOTCH2* locus, and determination of allele frequency distributions of Singly Unique Nucleotides (SUNs), base variants that occur only in one paralog (Sudmant et al., 2010), we established that the *NOTCH2NL* sequence emerged by a partial duplication of *NOTCH2*

prior to the last common ancestor (LCA) of human, chimpanzee and gorilla. Based on absence in orangutan and all other outgroups, it occurred after the LCA of human, chimpanzee, gorilla and orangutan (**Figure 2**). Both chimpanzee and gorilla have

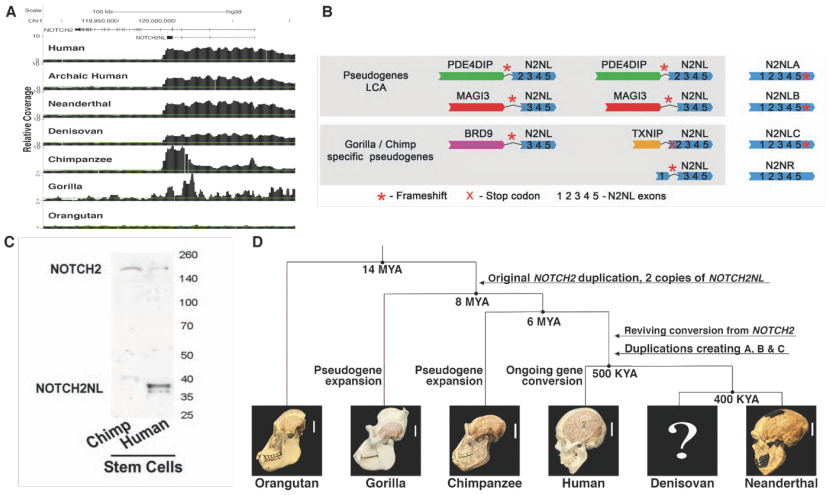


Figure 2. Evolutionary analysis of *NOTCH2NL*-like genes reveals only hominin *NOTCH2NL* genes encode NOTCH-related proteins. (A) Coverage of Illumina genome sequencing reads mapped to the *NOTCH2* locus demonstrating an excess of coverage in the 5' region of *NOTCH2* into intron 4 in all hominins examined, chimpanzee and gorilla, but not orangutan. (B) Schematic of NOTCH2NL-containing genes identified in gorilla, chimpanzee and human demonstrating that only in human *NOTCH2NL* genes encode NOTCH2-related proteins. (C) Immunoblot using an N-terminal NOTCH2 antibody (aa 25-255) of protein extracts from human and chimpanzee pluripotent stem cells showing expression of full-length NOTCH2 in both samples, but NOTCH2NL only in human. (D) Summary of the major events leading to the emergence of *NOTCH2NL*-related genes in the great ape lineage. See also Figure S2.

variable read depth over the region encompassing their *NOTCH2NL-like* sequences, suggesting they contain multiple versions of truncated *NOTCH2* putative pseudogenes sharing the same breakpoint within *NOTCH2* present in the human *NOTCH2NL* genes (**Figure 2A**). Using Bacterial Artificial Chromosome (BAC) clones and whole-genome shotgun contigs, we identified several individual *NOTCH2NL*-like genes in the chimpanzee genome, all of which were truncated and predicted to be non-functional

(**Figure 2B, Figure S2**). Some of these chimpanzee *NOTCH2NL*-like pseudogenes lack a 52 kb region containing exon 2. The resulting *NOTCH2NL*-like Δexon2 transcripts encode a protein of 88 amino acids with no homology to human NOTCH2NL (**Figure S2F)**, which was confirmed by ectopic expression of one of these chimp *NOTCH2NL*-like cDNAs **(Figure S2D).** From whole-genome shotgun contigs we found evidence for additional *NOTCH2NL*-like pseudogenes in the chimpanzee lacking either exon 1 or exon 1-2, located downstream of various 3' truncated genes (**Figure 2B, Figure S2B**) These make pseudogene fusion-transcripts of the form $PDE4DIP^{exon1\text{-}27}$-$NOTCH2NL^{exon2\text{-}5}$, $TXNIP^{exon1}$-$NOTCH2NL^{exon2\text{-}5}$ $MAGI3^{exon1}$-$NOTCH2NL^{exon3\text{-}5}$, and $MAGI3^{exon1\text{-}14}$-$NOTCH2NL^{exon3\text{-}5}$ as confirmed by RT-PCR using RNA derived from chimpanzee induced pluripotent stem cells (IPSCs) (**Figure S2C**). Sequence analysis of these transcripts established that the *NOTCH2NL* exons in all of these fusion-genes were out of frame with the upstream exons, indicating that these fusion genes cannot produce functional NOTCH2NL proteins (**Figure S2F)**. In support of this result, whereas we can detect NOTCH2 protein in extracts derived from chimpanzee IPSCs, we cannot detect a NOTCH2 N-terminal antibody reactivity in the molecular weight range expected for NOTCH2NL proteins under the same conditions where we can detect human NOTCH2NL protein (**Figure 2C**). Overall, the analysis of the configuration of *NOTCH2NL*-like genes in the chimpanzee genome and experimental evidence of lack of the protein indicates that chimpanzees lack a functional *NOTCH2NL* gene.

In the gorilla genome, three fusion pseudogenes involving *NOTCH2NL*-like sequence were identified (**Figure S2A**). In one gorilla BAC, exons 1-12 of *BRD9* are fused to exons 3-5 of NOTCH2NL. Surprisingly, two other gorilla fusion pseudogenes were

similar to *NOTCH2NL*-fusion pseudogenes found in chimpanzee: a *PDE4DIP*$^{exon1-27}$-*NOTCH2NL*$^{exon2-5}$ fusion was supported by BAC sequence and a *MAGI3*$^{exon1-14}$-*NOTCH2NL*$^{exon3-5}$ fusion was supported by trace archive sequence. Transcript support was obtained for *BRD9-NOTCH2NL* and *PDE4DIP*$^{exon1-27}$*-NOTCH2NL*$^{exon2-5}$ using RNA from gorilla iPSCs (Marchetto et al., 2013) (**Figure S2C)**.

The presence of two distinct fusion pseudogenes in both gorilla and chimpanzee, *PDE4DIP-NOTCH2NL* and *MAGI3-NOTCH2NL,* suggests that these two pseudogenes were established in the LCA of human, chimp and gorilla. Yet none of the chimpanzee or gorilla pseudogenes are found in the human genome. Further, human *NOTCH2NL* genes are all in the vicinity of *PDE4DIP*, but in stark contrast to the *PDE4DIP-NOTCH2NL* fusion genes in chimp and gorilla, human *NOTCH2NL* genes have a 5' genomic structure highly similar to *NOTCH2*. This suggests a plausible evolutionary history of *NOTCH2NL* genes as follows (**Figure 2D**): Both the *PDE4DIP-NOTCH2NL* and *MAGI3-NOTCH2NL* fusion pseudogenes were present the common ancestor of human, chimp, and gorilla between 8 and 14 MYA. Then, only in the human lineage, the ancestral *PDE4DIP-NOTCH2NL* fusion gene was 'revived' by *NOTCH2* through ectopic gene conversion. With the acquisition of exon 1 and the upstream promoter, this created a viable *NOTCH2NL* gene encoding a stable NOTCH-related protein. Because no remnants of the *MAGI3-NOTCH2NL* fusion pseudogene are found in the human genome, this must have been lost, perhaps in the upheaval of the pericentric inversion on chromosome 1 and subsequent large-scale copy number changes (Szamalek et al., 2006). The revived human *NOTCH2NL* subsequently duplicated twice more to form the cluster of three nearly identical *NOTCH2NL* genes on the q arm of chromosome 1 in the

human genome. Chimpanzee and gorilla had additional, species-specific *NOTCH2NL*-like gene duplications, but none produced functional genes.

It should be noted that ectopic gene conversion is not unusual in this region (Nuttle et al., 2013). This scenario involves quite a number of lineage-specific gene duplications, rearrangements and other losses. Close neighbors of *NOTCH2NL* have a similarly elevated number of lineage-specific structural events, e.g., the *NBPF* paralogs, which are estimated to have had six lineage-specific duplications in the human lineage (O'Bleness et al., 2014). This propensity for segmental duplication, loss, rearrangement and ectopic gene conversion is characteristic of what have been called duplication hubs in the ape genomes (Bailey and Eichler, 2006). *NOTCH2NL* is part of such a hub.

To explore when the gene conversion event that revived *NOTCH2NL* occurred, we calculated the number of substitutions per kilobase between *NOTCH2* and *NOTCH2NL*, and we analyzed three archaic human genomes (Lazaridis et al., 2014), two Neanderthal genomes (Prufer et al., 2017), (Prufer et al., 2014) and one Denisovan genome (Meyer et al., 2012). While archaic humans appear to have three *NOTCH2NL* loci (presumably A, B and C) and one *NOTCH2NLR* like modern humans, the Neanderthal and Denisovan genomes have only A, B and C-like sequence, and lack *NOTCH2NLR*. The presence of *NOTCH2NLA*, *B*, and *C* in all 3 hominin genomes suggests that the ectopic gene conversion creating hominin *NOTCH2NL* happened prior to their LCA, more than 0.5 MYA. The substitution rate between human *NOTCH2* and human paralogs *NOTCH2NLA*, *B* or *C* is roughly half of that between human and chimp, which, if calibrated to a human-chimp divergence of 6.5 MYA, gives a date for the ectopic gene conversion between 3 and 4 MYA. This corresponds to a time just before

or during the early stages of the expansion of the hominin neocortex (Holloway et al., 2004) (**Figure S2E**).

NOTCH2NL is expressed in Radial Glia neural stem cells during human cortical development

NOTCH2NL shares promoter sequence with *NOTCH2* and is therefore expected to be expressed in a pattern similar to *NOTCH2*. To learn what cell types in the developing brain express *NOTCH2NL*, we examined *NOTCH2NL* expression in 3,466 single cells derived from human fetal brains ranging in age from 11 to 21.5 post coital weeks (pcw) that were sampled from multiple regions of the dorsal and ventral telencephalon (Nowakowski, et al, *Science*, In Press, <bit.ly/cortexSingleCell>) (**Figure 3**).

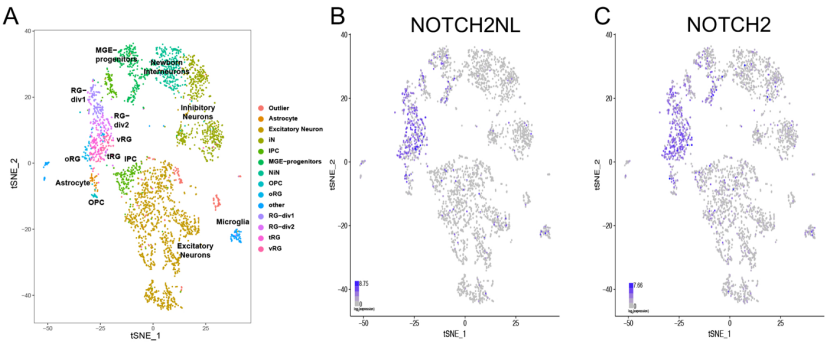


Figure 3. Radial glia-specific expression of *NOTCH2NL* in human fetal brain samples. Scatterplot of 3466 fetal brain cells after principal components analysis and t-stochastic neighbor embedding (tSNE) samples as described in Nowakowski, et al, 2017. Cells are colored by annotated cell type clusters (A), *NOTCH2NL* expression (B) and *NOTCH2* expression (C). See also Figure S3.

This data set contains a broad diversity of cell types including radial glia neural stem cells, intermediate progenitor cells (IPC), excitatory and inhibitory neurons, oligodendrocyte progenitors (OPC), astrocytes and microglia. The *NOTCH2NL*

expression pattern we found closely resembles that of *NOTCH2* and is highest in various radial glia populations, including outer radial glia cells (oRG), as well as astrocytes and microglia (**Figure 3, Figure S3**). *NOTCH2NL* expression in the oRG population is especially interesting as this pool of neural stem cells is thought to be responsible for much of the extended neurogenesis unique to human. This neurogenesis occurs in the more sparsely packed outer subventricular zone and recent transcriptional analysis suggests that oRG cells themselves express factors to support the subventricular niche (Pollen et al., 2015). In this setting, a secreted factor like NOTCH2NL might have a critical role in promoting maintenance of the neural stem cell fate.

We also estimated the relative expression of the H9 hESC paratypes of *NOTCH2NL* in bulk RNA undifferentiated cells and week 5 cortical organoids derived from H9 ESCs (see below). We used two different methods for this analysis. First, we measured the relative abundance using paratype-specific features in Illumina short-read RNA-seq data (**Figure S3B**). This analysis suggests that *NOTCH2NL*^{L} has the highest expression, and the other *NOTCH2NL* paratypes are expressed at levels 20-60% of *NOTCH2NL*^{L}. These estimates have some degree of uncertainty as much of the short read data is uninformative as to which paratype is being measured. As an alternative approach, we made a full-length cDNA library enriched for transcripts originating from chromosome 1q21.1 using biotinylated oligos and sequenced this library using a MinION nanopore sequencer. The resulting full-length *NOTCH2NL* transcripts were then assigned to their appropriate paratype, which allowed for a more accurate estimation of relative paratype expression (STAR Methods). The results confirmed that

NOTCH2NL^{L} has the highest expression, accounting for 43.6% of the transcripts, and indicated that the other paratypes have expression levels between ~8% and 40% of *NOTCH2NL*^{L} (**Figure 1C**).

Ectopic expression of NOTCH2NL delays mouse cortical neuron differentiation

To address the role of *NOTCH2NL* in cortical development, we assessed the effects of ectopic NOTCH2NL in mouse cortical organoids. We generated stable cell lines of mouse ESCs ectopically expressing human *NOTCH2NL*^{Sh,T197I}, and mESC EV control cell lines. mESCs were differentiated into cortical organoids using an established protocol (Eiraku et al., 2008) **(Figure 4A)**.

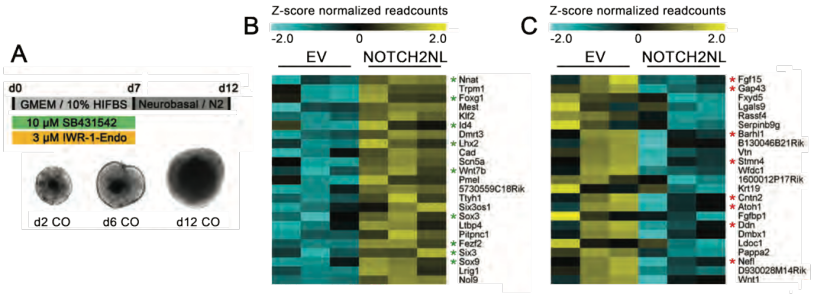


Figure 4. Ectopic expression of *NOTCH2NL* delays neuronal differentiation. (A) Overview of mouse cortical organoid differentiation protocol. (B) Genes upregulated in mouse organoids ectopically expressing *NOTCH2NL*^{Sh,T197I} compared to EV. Asterisks indicate genes associated with radial glia cells. (C) Genes downregulated in mouse organoids ectopically expressing *NOTCH2NL*^{Sh,T197I} compared to EV. Asterisks indicate genes involved in neuron differentiation. n = 3 per condition, pools of 16 organoids per replicate. See also Figure S4 and Table S2.

At day 6 of differentiation, a stage where most cells express the neural stem cell marker Pax6, mouse cortical organoids were harvested for RNA sequencing (RNA-seq). Differential expression analysis identified 147 differentially expressed genes (p-adj <

0.05, **Figure S4A-B, Table S2**). Mouse organoids ectopically expressing NOTCH2NL showed increased expression of genes involved in negative regulation of neuron differentiation, such as *Foxg1, Id4, Fezf2, Sox3* and *Six3* **(Figure 4B)** and several genes associated with neuronal differentiation were downregulated, including *Cntn2, Nefl, Gap43,* and *Sox10* **(Figure 4C, Figure S4C)**. These results suggest ectopic expression of NOTCH2NL in mouse organoids delays differentiation of neuronal progenitor cells.

Deletion of NOTCH2NL affects development of human cortical organoids

To explore the functional role of *NOTCH2NL* in human cortical development, we used the CRISPR/Cas9 system to generate genetic deletions of *NOTCH2NL* genes in hESCs. To avoid targeting *NOTCH2*, whose sequence is >99% identical to *NOTCH2NL* even in intronic regions, we used two guides, one in intron 1 with a 1 base mismatch

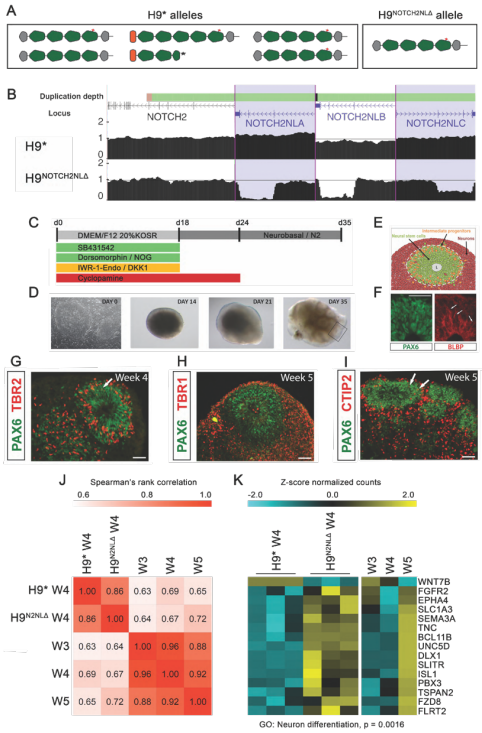


Figure 5. Cortical organoids from hESCs lacking *NOTCH2NL* show premature neuronal maturation. (A) Schematic showing *NOTCH2NL* alleles present in control (H9*) and NOTCH2NL mutant (H9$^{NOTCH2NL\Delta}$) cell lines. (B). Multi-region UCSC genome browser view of *NOTCH2* and *NOTCH2NL* genes with tracks showing normalized genome sequencing coverage demonstrating homozygous loss of exon 2-5 sequence for *NOTCH2NLA* and

NOTCH2NLB and heterozygous loss for *NOTCH2NLC*. Some coverage is seen in *NOTCH2NLA* and *NOTCH2NLB* due to a small portion of ambiguous linked read barcodes. Schematics of the cortical organoid protocol used (C) with pictures showing cells at various stages (D) and cell types generated. (F-I) Immunofluorescence staining of H9 hESC cortical organoids with markers of radial glia (PAX6, BLBP), intermediate progenitor cells (TBR2) and deep layer excitatory projection neurons (TBR1, CTIP2). (J) Spearman's rank correlation plot of the top 250 upregulated and downregulated genes ($H9^{NOTCH2NL\Delta}$ / H9*), and the matching data in W3, W4 and W5 H9 organoids. Numbers in plot indicate pairwise correlation values. (K) Heatmap showing expression profiles for a selection of genes in the significantly enriched GO cluster 'neuron differentiation'. n = 3.

with *NOTCH2* and *NOTCH2NLR*, but identical to the corresponding sequence in all H9 1q21 *NOTCH2NL* genes, and another that spans a 4 bp deletion relative to *NOTCH2* at the start of exon 5. This region is also quite different in *NOTCH2NLR* (13/20 mismatches to *NOTCH2NL*) (**Figure S5A**).

Clones were screened by PCR for the large deletion from intron 1 to exon 5 and then several were expanded and analyzed by targeted linked-read sequencing and assembly as described above to determine which NOTCH2NL genes were affected and to evaluate potential off-target effects at *NOTCH2*. For one clone, the sequence analysis revealed that *NOTCH2NLA* and *B* genes showed a homozygous deletion, and NOTCH2NLC a heterozygous deletion, leaving only one NOTCH2NLC ($NOTCH2NL^{Sh\text{-}2ntdel}$) intact. This clone is denoted $H9^{NOTCH2NL\Delta}$ **(Figure 5A-B**). As a control, another clone was selected that went through the same CRISPR/Cas9 transfection and selection process, but does not harbor any alterations around the guide sites in *NOTCH2NL*, *NOTCH2NLR* or *NOTCH2* loci (denoted H9*).

Neuronal tissues were generated from these clones by directed differentiation of human H9 hESC into cortical organoids **(Figure 5C-D)**. hESC-derived cortical organoids resemble early developmental stages of primate cortex development (Eiraku et al., 2008) (**Figure 5E-I**), displaying neural rosette structures that contain radially organized cortical RG cells giving rise to cortical neurons (**Figure 5E-F**) (Lancaster et al., 2013), (Pasca et al., 2015). Bulk RNA-seq transcriptome analysis of cortical organoids isolated

at weekly time points reveals that human ESC-derived cortical organoids display efficient and selective induction of dorsal forebrain marker genes, highly resembling the expression pattern during early stages (8-9 post conception weeks) of human dorsal forebrain development *in vivo* (**Figure S5B, Table S2**). H9$^{NOTCH2NL\Delta}$ organoids remained slightly smaller compared to control (H9*) organoids **(Figure S5C)**. At week 4 (w4), cortical organoids were harvested for RNA-seq analysis and differentially expressed (DE) genes between H9$^{NOTCH2NL\Delta}$ and H9* were discovered by DESeq2 analysis (Love et al., 2014). To investigate a potential shift in timing of cortical neuron differentiation between H9$^{NOTCH2NL\Delta}$ and H9* organoids, gene expression of the top 250 up- and down-regulated DE genes was correlated to the previously generated RNA-seq profiles of H9 w3, w4 and w5 cortical organoids (**Figure S5B, Table S2**). This analysis reveals that while differentially expressed genes in w4 H9* correlate best with w4 H9 cortical organoids, strikingly, w4 H9$^{NOTCH2NL\Delta}$ showed a better correlation with w5 H9 cortical organoids (**Figure 5J**). This indicates that w4 H9$^{NOTCH2NL\Delta}$ organoids prematurely express characteristics of older w5 organoids, suggesting H9$^{NOTCH2NL\Delta}$ are advanced in their development compared to H9* organoids. To obtain better insights into what aspects of cortical organoid development were advanced by NOTCH2NL deficiency, we performed GO-term enrichment analysis for the selection of DE genes in H9$^{NOTCH2NL\Delta}$ which were found to correlate better with w5 H9 cortical organoids than organoids of w4 **(Figure 5K)**. This cluster of 212 DE genes was found to be enriched for genes involved in neuron differentiation, including key regulators of neuron differentiation such as *DLX1*, *BCL11B*, *SEMA3A, UNC5D* and *FGFR2*. These data suggest that w4

H9$^{NOTCH2NL\Delta}$ display premature differentiation, which further supports a potential role for NOTCH2NL in delaying differentiation of neuronal progenitors in the human cortex.

NOTCH2NL amplifies Notch-signaling through direct interaction with NOTCH receptors

We next sought to test whether *NOTCH2NL* can influence NOTCH2 signaling. The 6 N-terminal EGF domains encoded in NOTCH2NL are also present in the full-length NOTCH2 receptor. These specific domains do not have a clearly described function in activation of the NOTCH pathway. However, they are conserved from *Drosophila* to human, indicating that they are important for normal NOTCH receptor function. There is evidence that the N-terminal EGF domains are involved in

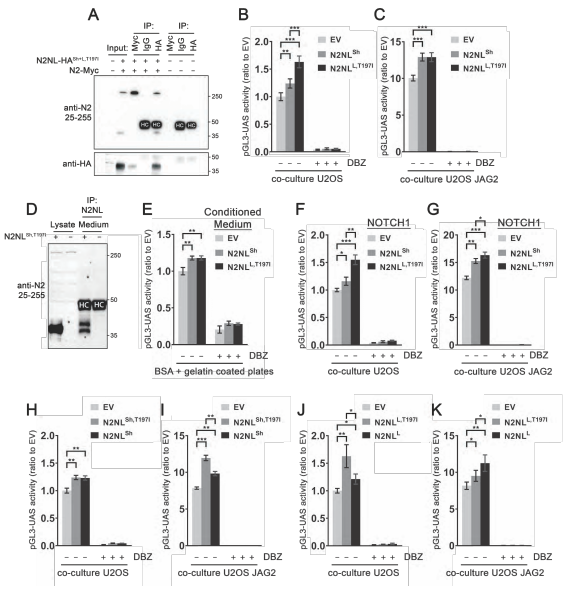


Figure 6. NOTCH2NL paratypes interact with NOTCH receptors and amplify NOTCH signaling. (A) Co-immunoprecipitation of NOTCH2 and NOTCH2NL analyzed by immunoblot. (B-C) Co-transfection of NOTCH2-GAL4 and NOTCH2NL boosts activity of the pGL3-UAS reporter. Upon stimulation of NOTCH2-GAL4 with JAG2 in a co-culture setup, NOTCH2NL remains effective in boosting reporter activity. Average of 4 independent experiments with 6 replicates each. Two-way ANOVA with Tukey's HSD (** p <10^{-8}, *** p <10^{-12}), error bars indicate SEM. (D) NOTCH2NL is secreted and can be immunoprecipitated from the medium of mouse ESCs ectopically expressing

NOTCH2NL. (E) Addition of NOTCH2NL-conditioned medium to cells transfected with NOTCH2-GAL4 and the pGL3-UAS reporter also enhances reporter activity. Average of 2 independent experiments with 4 and 3 replicates each. One-way ANOVA with Tukey's HSD (* $p < 10^{-4}$), error bars indicate SEM. (F-G) The effect of NOTCH2NL is not restricted to NOTCH2, in an assay with NOTCH1-GAL4 we observe similar effects. (H-K) The presence or absence of the ancestral start codon, and the amino acid 197 Thr→Ile variant, each show specific characteristics in the reporter assays. (H-I) Average of 2 independent experiments with 6 replicates each. Two-way ANOVA with Tukey's HSD (** $p < 10^{-8}$, *** $p < 10^{-12}$), error bars indicate SEM. (J-K) 6 replicates in one experiment. Student's t-test with Holm-Bonferroni correction (* $p < 0.05$, ** $p < 10^{-3}$), error bars indicate SD. See also Figure S6.

dimerization of NOTCH receptors, and that receptor dimerization modulates NOTCH activity (Duering et al., 2011), (Kopan and Ilagan, 2009), (Xu et al., 2015), (Nichols et al., 2007). For example, mutations in the N-terminal EGF domains of NOTCH3 are commonly found in patients suffering from a disease called CADASIL, a hereditary stroke disorder, (Joutel et al., 1997) and these mutations cause aberrant aggregation and dysfunction of NOTCH3 receptors (Karlstrom et al., 2002).

To establish a functional relationship between NOTCH2NL and NOTCH2, we searched for interactions between these proteins. By co-transfecting HA-tagged versions of *NOTCH2NL*^{Sh} and *NOTCH2NL*^{L,T197I} with a myc-tagged version of NOTCH2 in HEK293 cells we were able to detect co-immunoprecipitation of both proteins by pull-down of either NOTCH2NL or NOTCH2 **(Figure 6A)**.

Under these co-IP conditions, we did not find detectable interactions with the other EGF-like domain interacting proteins EGFR and PDGFRB **(Figure S6A-B)**. These findings indicate NOTCH2NL may influence the conformation and dimerization of NOTCH2.

To assess the influence of NOTCH2NL on activation of the NOTCH2 receptor, a luciferase reporter assay was used. The NOTCH2-ICD DNA binding domain was replaced with a GAL4 domain (NOTCH2-GAL4), and used together with a pGL3-UAS luciferase reporter. CMV-Renilla was used in all reporter experiments for normalization.

This allows for precise measurements of NOTCH2-GAL4 activation, without secondary effects from other NOTCH receptors or other pathways (Habets et al., 2015), (Groot et al., 2014). Co-transfection of empty vector (EV), *NOTCH2NL*^{Sh} or *NOTCH2NL*^{L,T197I} with *NOTCH2-GAL4* in U2OS cells increased pGL3-UAS reporter activity **(Figure 6B)**. *NOTCH2NL*^{Sh} increased activity by 24% and *NOTCH2NL*^{L,T197I} increased activity by 63%, indicating both the short and long forms of NOTCH2NL can enhance NOTCH receptor activation.

Next, we tested whether NOTCH2NL is still functional under active NOTCH signaling conditions. For this, U2OS cells were transfected with *NOTCH2-GAL4* and either EV, *NOTCH2NL*^{Sh} or *NOTCH2NL*^{L,T197I}. The transfected cells were then co-cultured on a layer of cells expressing the NOTCH ligand JAG (U2OS-JAG2-Myc) or control (U2OS) cells. Trans-interaction of NOTCH2-GAL4 with JAG2 leads to high activation of the pGL3-UAS reporter in this system. NOTCH2NL was able to increase cleavage of the NOTCH2-GAL4 receptor, indicating that even under these high-signaling conditions NOTCH2NL can further amplify Notch signaling **(Figure 6C)**.

We then tested interactions specific to secreted NOTCH2NL. To make secreted NOTCH2NL, mouse ESCs were transfected with *NOTCH2NL*^{Sh} or EV as a control. The medium was collected after 32 hours, and used for immunoprecipitation with NOTCH2 aa 25-255 antibody. The isolated protein samples were analyzed by Western blot, which confirmed the presence of secreted NOTCH2NL in the medium **(Figure 6D)**. The two bands of *NOTCH2NL*^{Sh} may represent the glycosylated protein (higher band) and unmodified protein (lower band). This pattern was also observed in ectopic expression of N-terminal fragments of the NOTCH3 receptor (Duering et al., 2011).

To determine if NOTCH2NL can act cell non-autonomously on NOTCH signaling, we generated medium conditioned with secreted NOTCH2NL or EV-transfected control cells. U2OS cells were transfected with only NOTCH2-GAL4 and the luciferase reporter. After 6 hours, cells were replated in either EV-, *NOTCH2NL*^{Sh}- or *NOTCH2NL*^{L,T197I}-conditioned medium harvested from other transfected U2OS cells. After 24 hours, cells were isolated for luciferase measurements. NOTCH2NL^{Sh}-conditioned medium increased pGL3-UAS reporter activity by 24% and NOTCH2NL^{L,T197I}-conditioned medium increased reporter activity by 22% **(Figure 6E)**. NOTCH2NL^{Sh} and NOTCH2NL^{L,T197I} have a different N-terminal structure **(Figure 1C)**, whereby NOTCH2NL^{Sh} lacks the NOTCH2 signal peptide for canonical transport through the secretory pathway, yet NOTCH2NL^{Sh} is equally potent as NOTCH2NL^{L,T197I} when present in conditioned medium. This suggests NOTCH2NL^{Sh} acts in a cell non-autonomous manner on Notch signaling after being secreted by an unconventional pathway (Steringer et al., 2015), (Rabouille, 2017). In contrast, in the co-transfection experiments NOTCH2NL^{L,T197I} was clearly more potent than NOTCH2NL^{L,T197I}-conditioned medium (**Figure 6B**). This suggests NOTCH2NL^{L,T197I} can amplify Notch signaling through both cell-non-autonomous and cell-autonomous mechanisms.

We next tested if NOTCH2NL could also influence the activity of the other NOTCH receptor paralogs NOTCH1 and NOTCH3. NOTCH2NL^{Sh} and NOTCH2NL^{L,T197I} were able to activate cleavage of both NOTCH1-GAL4 and NOTCH3-GAL4, both in co-culture with U2OS control cells and U2OS-JAG2 cells, at a level similar to NOTCH2-GAL4 **(Figure 6F-G, Figure S6C)**. To assess whether NOTCH2NL functions differently under various ligand stimulated conditions, we tested the ability of NOTCH2NL to

enhance NOTCH receptor activity in OP9 cells stably expressing DLL1 and control OP9 cells. NOTCH2NL^{Sh} and NOTCH2NL^{L,T197I} mediated activation of the NOTCH2-GAL4 receptor was essentially the same when stimulated by JAG2 or DLL1 **(Figure S6D)**. Lastly, NOTCH2NL was similarly able to activate NOTCH2-GAL4 in U2OS cells seeded on recombinant DLL4 coated plates **(Figure S6E)**. These data show that NOTCH2NL's amplifying effect on Notch signaling is not restricted to NOTCH2, and that the potency of NOTCH2NL to enhance NOTCH receptors is not specific for a particular NOTCH-ligand.

Unlike NOTCH2NL^{Sh}, NOTCH2NL^{L,T197I} harbors a common variant present in EGF domain 5 that changes the ancestral threonine at position 197 to isoleucine **(Figure 1C)**. This threonine can be fucosylated, while isoleucine cannot. Fucosylation is essential for modifying binding of EGF domains to their interaction sites. We tested the short and long versions of NOTCH2NL with and without the T197I variant in the NOTCH2-GAL4 - pGL3-UAS reporter assay. Interestingly, the single amino acid changes had subtle but significant influences on NOTCH2-GAL4 cleavage**.** *NOTCH2NL^{Sh}* and *NOTCH2NL^{Sh,T197I}* performed equally under baseline conditions (29% and 31% increase, respectively, **Figure 6H**). However, with JAG2 co-culture, *NOTCH2NL^{Sh,T197I}* variant was more potent than *NOTCH2NL^{Sh}* (53% and 25% increase respectively, **Figure 6I**). *NOTCH2NL^{L,T197I}* was also a more potent activator than *NOTCH2NL^{L}* (63% and 21% increase, respectively, **Figure 6J**). However, co-culture with JAG2 cells showed the opposite. Here, *NOTCH2NL^{L,T197I}* had less effect than *NOTCH2NL^{L}* (17% and 38% increase, respectively, **Figure 6K**).

Altogether, these data suggest that NOTCH2NL can enhance NOTCH receptor activation. Structural changes between the various NOTCH2NL paratypes, varying from the absence/presence of the signal peptide to single amino acid changes in the EGF-like repeats, show differential effects under different conditions. This implies that the influence of NOTCH2NL genes on cortical neurogenesis in any particular individual may be dependent on the combination of NOTCH2NL paratypes he or she carries. However, widespread genomic variation in NOTCH2NL paratypes in the healthy human population shows that the absence or loss of one specific paratype is not necessarily associated with disease, so there must be some robustness to the system as well.

Deletions and duplications of NOTCH2NL genes are associated with neurodevelopmental phenotypes

The revision of the sequence of the 1q21.1 band in the human reference genome GRCh38 for the first time clearly separated the Thrombocytopenia Absent Radius (TAR) syndrome's centromerically proximal locus from the 1q21.1 deletion/duplication syndrome's distal locus (Figure 1A). The latter is bracketed by *NOTCH2NLA* at GRCh38.chr1:146,151,907-146,229,032 (within a remapped region that was previously called breakpoint 3) and *NOTCH2NLB* at GRCh38.chr1:148,602,849-148,679,774 (aka breakpoint 4) (Mefford et al., 2008), (Rosenfeld et al., 2012) (**Figure S7**).

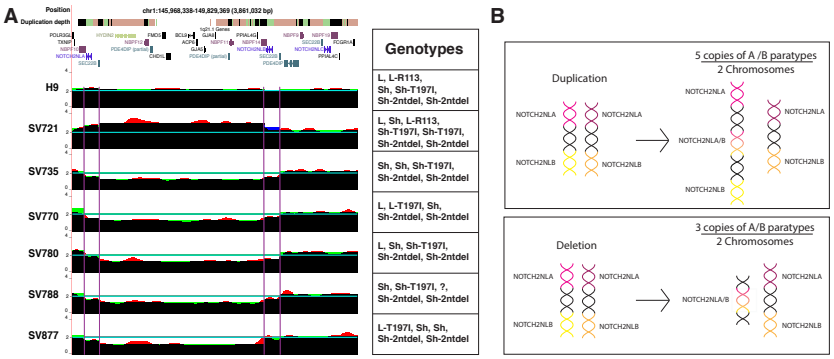


Figure 7. Patients with 1q21.1 Distal Deletion/Duplication Syndrome show breakpoints and CNV in *NOTCH2NLA* and *NOTCH2NLB*. (A) UCSC Genome Browser screenshot from GRCh38. The duplication depth track indicates duplicated genome sequences as colored bars: white (single copy, N=1), orange (N=2-4), green (N=5), black (N > 5). The normalized read depth for each individual is in red (see methods). For breakpoints in *NOTCH2NL*, due to sequence similarity and gene conversion, we expect the average normalized coverage of *NOTCH2NLA* and *NOTCH2NLB* to be 1.5 for deletions and 2.5 for duplications. Similarly, we expect the normalized coverage of the sequence in between to be 1.0 for deletions and 3.0 for duplications (blue). The region was then broken into 5 portions -- centromeric to *NOTCH2NLA*, *NOTCH2NLA*, in between *NOTCH2NLA* and *NOTCH2NLB*, *NOTCH2NLB* and telomeric to *NOTCH2NLB*. The average values of the normalized coverage based on the five regions are drawn in green. Where all models agree, the colors combine to black. (B) Schematic of the *NOTCH2NL* chromosomal configuration before and after a duplication or deletion event. In the case of duplication, there is an extra NOTCH2NL paratype that is a hybrid of the *NOTCH2NL* paratype at the *NOTCH2NLA* position and that at the *NOTCH2NLB* position from that position. For a deletion, only a hybrid of the two *NOTCH2NL* paratypes from that chromosome remains. See also Figure S7, Table S3-S4.

These breakpoints are 99.7% identical over an interval of approximately 250,000 bases, creating frequent non-allelic homologous recombination that results in either deletion or duplication of a 2.4 MB segment in the chromosome in which it occurs. Supporting the high frequency of the event, it is estimated that 18-50% of distal duplications and deletions seen in the clinic are *de novo* (Haldeman-Englert and Jewett, 1993). The remainder inherit the condition, often from an asymptomatic or mildly symptomatic parent. The frequency of the deletion in the European general population has been estimated at 0.03% and the duplication at 0.049% (Mace et al., 2017).

When present, symptoms are strikingly, but not exclusively, neurological. In several independent large studies 1q21.1 distal deletion has been shown to be associated with schizophrenia (International Schizophrenia Consortium, 2008; Mace et al., 2017), (Grozeva et al., 2012), (Chang et al., 2016). In contrast, in patients ascertained as part of the Simons Variation in Individuals (VIP) autism study, 1q21.1 distal duplication probands exhibited ADHD (29%), behavior disorder (18%), autism spectrum disorder (41%), developmental coordination disorder (23%), intellectual disability (29%), while deletion probands exhibited these symptoms at lower frequencies, but with a relatively high percentage (26%) exhibiting anxiety and mood disorders (Bernier et al., 2016). Of particular interest regarding the function of NOTCH2NL in the enhancement of NOTCH signaling during neurodevelopment is the fact that studies have consistently shown that the duplication is associated with macrocephaly and the deletion with microcephaly (Mefford et al., 2008), (Brunetti-Pierri et al., 2008).

Because the copy number variation studies that identified the 1q21.1 syndromes were performed on microarrays that were based on the GRCh37 and earlier assemblies that were incorrect in the 1q21.1 band, we reanalyzed copy number variation microarray data derived from 11 previously characterized patients where remappable data could be obtained, 9 with microcephaly and 2 with macrocephaly, to determine the boundaries of their 1q21.1 variations in GRCh38 coordinates. Probe positions were realigned to the revised assembly, and a more sensitive method of copy number variation analysis involving optimization by integer linear programming was developed (STAR Methods). The remapped data suggest that *NOTCH2NLA* and *NOTCH2NLB* are located inside the 1q21.1 locus implicated in microcephaly and macrocephaly (**Figure S7**). Nine out of the

nine microcephalies were consistent with *NOTCH2NLA* and/or *-B* deletion and both macrocephalies were consistent with *NOTCH2NLA* and/or *-B* duplication. In at least one out of nine microcephaly patients (number 2), the *HYDIN2* locus exhibits a normal copy number, consistent with results previously published (Dougherty et al., 2017).

The sparsity of unique probes in the highly repetitive 1q21.1 region precludes a precise breakpoint analysis of these older data. In order to better map the breakpoints of the copy number variations in 1q21.1 patients, we obtained primary fibroblasts from six autism patients in the Simons VIP project (Bernier et al., 2016) one listed as having 1q21.1 distal duplication and five identified as having the distal deletion (**Figure 7**, **Table S3**), and prepared high molecular weight DNA for targeted sequencing and assembly as described above. The sequencing coverage data demonstrates that all of these samples have their deletion or duplication in the breakpoint 3 and 4 regions containing *NOTCH2NLA* and *NOTCH2NLB* (**Figure 7A**). In particular, in deletions we see normal copy number 2 for unique DNA outside of the locus and copy number 1 for unique DNA inside the locus. In the two near-identical breakpoint regions that flank the locus we expect both NOTCH2NLA and NOTCH2NLB to be present on the non-deleted chromosome, and a homology-driven deletion in the other chromosome is expected to leave a single NOTCH2NL A/B gene, possibly a fusion gene (**Figure 7B**). Indeed, we see combined read coverage 3 for both breakpoint regions, which is split into two, leaving an “intermediate level plateau” of read depth at roughly 1.5 in each of the NOTCH2NL regions. In the case of a duplication, these intermediate level plateaus at the breakpoint flanks are at read depth 2.5, implying 5 copies of *NOTCH2NLA* and *B*.

When we assembled the *NOTCH2NL* paratypes in these individuals we confirmed that indeed all of the deletion samples have three *NOTCH2NLA* or *NOTCH2NLB* alleles and the duplication sample has five. In all but one sample (SV788) we could determine all of the *NOTCH2NLA/B* paratypes as well (**Figure 7, Table S1**). As expected from the position of the breakpoints in the coverage analysis, all 6 individuals showed no evidence of gain or loss of *NOTCH2NLC*, which has two *NOTCH2NL*$^{Sh\text{-}2ntdel}$ alleles. The patients, like normal individuals, show a variety of combinations of *NOTCH2NL* paratypes. If there is any pattern in the set of *NOTCH2NL* paratypes that is particular to 1q21.1 distal deletion/duplication syndrome, this patient set is too small to discern it. Nevertheless, this analysis suggests that the typical 1q21.1 distal deletion/duplication syndrome events are driven by non-allelic homologous recombination between *NOTCH2NLA* and *B* or nearby sequence, and these events produce *NOTCH2NL* copy number change. Thus, it remains possible that *NOTCH2NL* copies not only provide the breakpoints for these events, but may have a contribution to the neurodevelopmental phenotypes associated with them as well.

Discussion

NOTCH2NL joins *ARHGAP11B* (Florio et al., 2015) and *BOLA2* (Nuttle et al., 2016) as a third example of the concomitant emergence of a potential adaptive evolutionary innovation and susceptibility to a recurrent genomic disorder from chromosomal instability mediated by hominin-specific duplication hubs. These duplicated regions not only represent the bulk of human-specific genomic DNA (Cheng et al., 2005), but are an important a reservoir of new genes, carrying with them the potential to rapidly change the structure of the genome via non-allelic homologous recombination. Indeed the most

promising genes recently implicated in human brain evolution in addition to those above, such as *SRGAP2C* (Dennis et al., 2012), (Charrier et al., 2012) and *TBC1D3* (Ju et al., 2016) map to regions of human-specific structural variation.

The peculiar evolutionary history of *NOTCH2NL* includes a series of complex genomic reorganization events that eventually led to the creation of three functional NOTCH-related genes early in hominin evolution. *NOTCH2NL*-like pseudogenes lacking coding potential, including a *PDE4DIP-NOTCH2NL* pseudogene, were already present in the gorilla, chimpanzee and human LCA and still remain in the genomes of chimpanzee and gorilla, but are not found in hominin genomes. Instead, in each of the three hominin species sequenced, human, Neanderthal and Denisovan, we find 3 *NOTCH2NL* genes, *NOTCH2NLA*, *B* and *C*, and in humans we find an additional pseudogene, *NOTCH2NLR*. The most plausible explanation for this is that in a common ancestor of humans, Neanderthals and Denisovans, the ancestral *PDE4DIP-NOTCH2NL* pseudogene was repaired by ectopic gene conversion from *NOTCH2*. This event may have been crucial to hominin evolution, as it marks the birth of a novel hominin-specific NOTCH-related gene that likely functions to delay yet eventually increase production of neurons from neural stem cells during fetal brain development, and whose introduction was prior to or during a time of rapid expansion of brain size in hominins approximately 3 MYA (Halloway, et al., 2004) (**Fig. S2E**). Subsequent duplication events created additional *NOTCH2NL* genes in hominin ancestors, increasing *NOTCH2NL* dosage. In modern humans, three *NOTCH2NL* genes are present with at least eight different “paratypes”, i.e., haplotype configurations that can occur in different paralogous locations.

Consistent with a potential role in brain development, *NOTCH2NL* is expressed in the germinal zones of the developing human cortex where it likely promotes NOTCH signaling. Of particular interest is the expression of NOTCH2NL in outer radial glia cells, which are responsible for a more pronounced proliferative zone in the human developing cortex compared to that in related primates and are thought to be involved in the evolutionary expansion of the human cortex (Hansen et al., 2010, Lui et al., 2011). It has remained elusive how these loosely organized oRG cells maintain their sustained capacity for self-renewal and proliferation, as they lie in the outer subventricular zone (OSVZ) far away from the proliferative zone of strictly laminar radial glial cells near the surface of the ventricle. Recent work suggests that oRG cells directly contribute to a stem cell niche of the OSVZ through increased expression of extracellular matrix proteins and growth signal factors (Pollen et al., 2015). It has been shown that Notch signaling is essential for oRG cell self-renewal (Hansen et al., 2010), and our data suggest NOTCH2NL may act as a potentiating factor on Notch signaling to support oRG cell self-renewal. In particular, we show that NOTCH2NL can act in part in a cell non-autonomous manner to amplify Notch signaling. Our data also supports a separate, cell-autonomous function of NOTCH2NL, and indeed, in a very recent related study (Suzuki, et al, Submitted), it was found that the long version of NOTCH2NL ($NOTCH2NL^{L,T197L}$) promotes NOTCH signaling and neural progenitor proliferation in a cell-autonomous manner. Through these actions NOTCH2NL could support self-renewal of oRG cells in the OSVZ environment.

Enhancement of NOTCH signaling is likely the mechanism that underlies the gene expression phenotypes we observed in experiments in which NOTCH2NL is ectopically

expressed in mouse or is deleted in human cortical organoids. We showed that stable expression of NOTCH2NL in mouse organoids leads to downregulation of genes involved in neuronal differentiation and upregulation of genes involved in negative regulation of neuronal progenitor differentiation. These results are further supported in human cortical organoids grown from hESC lines with five out of six alleles of *NOTCH2NLA/B/C* deleted by CRISPR/Cas9, where NOTCH2NL deficiency resulted in premature expression of genes involved in neuronal differentiation.

The combined role of NOTCH2NL genes in any particular genotype may depend on the combined expression pattern of structurally distinct NOTCH2NL paratypes. The short forms of NOTCH2NL may act primarily in a cell non-autonomous way, as they are likely secreted by unconventional secretion pathways due to the lack of a signal peptide. Several routes of unconventional protein secretion have been described (Rabouille, 2017), of which FGF2 secretion is a well-studied example (Steringer et al., 2015). In contrast, the long forms of NOTCH2NL contain the signal peptide required to enter the canonical secretory pathway, where they can encounter NOTCH receptors and NOTCH ligands in secretory vesicles before they are transported to the outer membrane of the cell. The long forms of NOTCH2NL therefore may act primarily cell autonomously or may act both autonomously and non-autonomously. We found that the different paratypes of NOTCH2NL have different potencies to amplify NOTCH activity and do so under different circumstances (presence or absence of Notch ligand, secreted in the medium or generated within the cell, etc.). This suggests there may have been an evolutionary gene-dosage optimization, where NOTCH2NL paratypes with different

structural and sequence properties have undergone multiple changes to reach a balanced state of NOTCH pathway modulation.

The existence of many distinctly functional NOTCH2NL paratypes, created by substitutions and small indels that were likely to have been combinatorially amplified by ectopic gene conversion during hominin evolution, is striking. While this has resulted in many versions of NOTCH2NL encoded at the *NOTCH2NLA* and *NOTCH2NLB* positions in chromosome 1q21, the overall copy number of *NOTCH2NLA and NOTCH2NLB* is remarkably stable at a combined copy number 4 in almost all the 266 individuals of the Simons Diversity collection, with some being ambiguous. This suggests total dosage of these A/B paratypes may be important. In contrast, *NOTCH2NLC* has distinct paratypes (characterized primarily by a 2 bp deletion that forces use of a downstream start codon, denoted $NOTCH2NL^{Sh\text{-}2ntdel}$), and does not show as stable a copy number in the Simons collection (appearing to have copy number 3 in 3 individuals, copy number 1 in 20 individuals, and 0 in 1 individual). Further, despite the apparent conservation of A/B copy number, one of the A/B paratypes, $NOTCH2NL^{LR113*}$, contains a premature stop codon that is predicted to render the protein nonfunctional. This paratype appears to be present at copy number 2 in 4 individuals and copy number 1 in 31 in the Simons collection. There also appear to be 4 individuals who have 1 copy of $NOTCH2NL^{LR113*}$ and lack one copy of $NOTCH2NL^{Sh\text{-}2ntdel}$, so copies of $NOTCH2NL^{Sh\text{-}2ntdel}$ in the C paralog position are not always there to make up for stop codon-containing copies in the A or B positions.

Detailed modeling of selective forces on NOTCH2NL is complicated by the relatively small number of individuals whose genotypes have been resolved into a specific

combination of paratypes, and the peculiar mix of substitution and ectopic gene conversion between equivalent loci that creates a diversity of paratypes. Lacking large scale sequencing with long DNA fragments, the selective forces on *NOTCH2NL* genotypes remain mysterious. Still, it is clear that *NOTCH2NL* genotypes exist with different dosages of different functionally distinct paratypes in the population. Either the physiological system is quite robust to these protein differences despite their differential effect on Notch signaling, or the optimal balance of *NOTCH2NL* alleles has not been fixed in the population.

Despite their presence in the breakpoints of the 1q21.1 distal deletion/duplication region, *NOTCH2NL* genes have not previously been associated with this syndrome because the reference genome assembly was incorrect until the GRCh38 assembly. By reanalyzing older data and generating new data from patients from whom long genomic DNA fragments could be isolated, we find that all patients we have examined have copy number variations that include an extra *NOTCH2NL* copy in the duplication and loss of a copy in the deletion. Further, wherever it was possible to determine, we find that non-allelic homologous recombination has occurred inside or near *NOTCH2NLA* and *B*, leaving behind an intact *NOTCH2NL* copy, possibly a chimera of A and B (**Figure 7B**). Therefore, *NOTCH2NLA* and *B* should be considered candidates for contributing to the phenotypes of 1q21.1 deletion/duplication syndrome, including abnormal brain size phenotypes (Mefford et al., 2008), (Brunetti-Pierri et al., 2008), (Dumas et al., 2012), (Girirajan et al., 2013). However, they are not the only genes whose copy number is changed. Protein coding genes between *NOTCH2NLA* and *NOTCH2NLB* include *HYDIN2, CHD1L, BCL9, GJA5, GJA8, PRKAB2, FMO5, ACP6, PPIAL4G, GPR89B*,

and DUF1220-domain containing *NBPF11, NBPF12 and NBPF14* paralogs, three of the more than fourteen functional paralogs of *NBPF*. The progressive increase in number of DUF1220 domains in primate genomes has been shown to correlate with the evolutionary expansion of the neocortex (Popesco et al., 2006), and *CHD1L* is amplified in many solid tumors and promotes tumor growth (Cheng et al., 2013), (Xu et al., 2013), thus it could be associated with cell proliferation. *HYDIN2* was the strongest candidate for brain size effects of 1q21.1 distal deletion/duplication syndrome as it shares a protein domain with the microcephaly-associated protein *ASPM* and its paralog *HYDIN* lies in the 16q22.2 locus that has also been associated with microcephaly. However, recent work found 6 symptomatic patients with atypical breakpoints excluding *HYDIN2*, thus eliminating it as a likely driver (Dougherty et al., 2017). Mapping of these atypical patients indicates that the *NOTCH2NLA* locus was also excluded in these atypical events, but the *NOTCH2NLB* locus was fully duplicated in the three atypical duplication cases (all macrocephalic), and fully deleted in the three atypical deletion cases (all microcephalic). Therefore, these cases are consistent with a possible contribution of NOTCH2NL to 1q21.1 deletion/duplication syndrome. There are scattered reports of patients with smaller duplications or deletions, a few of these including neither *NOTCH2NLA* nor *NOTCH2NLB* (Girirajan et al., 2013), (Van Dijck et al., 2015). These studies have variously pointed to *CHD1L*, *BCL9* or the noncoding RNA gene *LINC00624* as candidates for causing 1q21.1 deletion/duplication syndrome symptoms. However, no consistent pattern has been found and no mechanistic explanation for the possible association of these genes with neurodevelopmental changes has yet been established. 2.4 megabases is a considerable stretch of DNA, and it may well be that

these or other genes, such as the three *NBPF* paralogs in the region, contribute to the neurological effects of this large copy number change. Indeed non-neurological 1q21.1 distal deletion/duplication effects in the heart and eyes are thought to be caused by the copy number changes in *GJA5* and *GJA8* (Gollob et al., 2006) (Shiels et al., 1998), so it is plausible that several genes may contribute to the apparently complex neurological effects. At the very least, by providing the substrates for non-allelic homologous recombination, *NOTCH2NL* genes enable the 1q21.1 distal deletion/duplication syndrome.

The strongly directional association of 1q21.1 distal deletion/duplication syndrome with brain size, with duplications tending to cause macrocephaly and deletions microcephaly, is most intriguing from a NOTCH2NL perspective, given that duplications are associated with increased NOTCH2NL dosage, which in *ex vivo* experiments delays differentiation of neuronal progenitor cells allowing for a longer period of proliferation, and deletions decrease NOTCH2NL dosage, which promotes premature differentiation of neuronal progenitor cells. While still circumstantial, these data support a role of *NOTCH2NL* genes not only in the cognitive and brain size symptoms of 1q21.1 distal deletion/duplication syndrome, but in the evolution of the hominin-specific larger brain size and associated cognitive traits. A delay in cortical maturation, coupled with a net increase in the size of the neocortex, in large part from sustained activity of outer radial glia progenitor cells, is characteristic of human brain development. Humans may in fact be caught in an evolutionary compromise in which having multiple identical copies of *NOTCH2NL* provides a neurodevelopmental function we need while at the same time predisposing our species to recurrent *de novo* non-allelic homologous recombination

events that underlie a neurodevelopmental syndrome and contribute to our overall genetic load. Given the many different alleles ("paratypes") of *NOTCH2NL* we observe segregating in the current human population, the tension caused by this compromise may still be a factor in our ongoing evolution.

Acknowledgements

We thank Robert Kuhn for help in obtaining data from researchers in the ISCA and DECIPHER consortia, and Mark Diekhans, Brain Raney, and Hiram Clawson for help with features of the UCSC Genome Browser used in this research; Raphael Bernier for help interpreting the Simons VIP phenotype data; Mari Olsen (Haussler lab), Bari Nazario (the Institute for the Biology of Stem Cells), Nader Pourmand (the UCSC Genome Sequencing Center), Ben Abrams (the UCSC Life Science Microscopy Center) and Shana McDevitt (UC Berkeley, QB3, Vincent J. Coates Genomics Sequencing Laboratory) for their excellent technical support; Arjan Groot and Marc Vooijs for generous supply of reagents and support for reporter assays; Martijs Jonker (SILS, University of Amsterdam) for help in RNA-seq data processing; Pierre Vanderhaeghen, Bin Chen, Benedict Paten, Ed Green and members of the Haussler and Jacobs labs provided helpful discussions and comments on the manuscript. This work was supported by CIRM Predoctoral Fellowship T3-00006 (ARF), CIRM Postdoctoral Fellowship TG2-01157 (FMJJ), Human Frontier Science Program Postdoctoral Fellowship LT000689/2010-L and HFSP Career Development Award CDA00030/2016C (FMJJ), Marie Curie reintegration grant H2020-MSCA-IF-2014_RI 659193 (FMJJ), NWO Earth and Life Sciences (ALW), project 834.12.003., EMBO ALTF 292-2011 (MH) and a special fellowship from donor Edward Schulak (ADE), NIH HG002385 (EEE), NIH

F30HG009478 (MLD), CIRM Center of Excellence for Stem Cell Genomics (Stanford) GC1R-06673-A, CIRM Center for Big Data in Translational Genomics (SALK) GC1R-06673-B, and StemPath NIH/NIGMS R01 GM109031 (DH), NCI Cloud award fund #24074-443720 (DH), and the California Institute for Quantitative Biosciences. David Haussler and Evan E Eichler are Investigators of the Howard Hughes Medical Institute.

Author Contributions

Conceptualization- D.H., S.R.S., F.M.J.J., I.T.F.; Methodology- I.T.F., S.R.S., D.H., M.M., A.D.E., G.L.M., A.Bi., A.M.N., F.M.J.J.; Validation- M.J.D., X.N.; Investigation- G.A.L., M.M., G.L.M., Avd.B., J.L.R., A.R.F., L.R., T.J.N., A.A.P.; Formal Analysis- I.T.F., A.D.E., A.M.N., A.B., R.L.R., A.B., M.L.D., X.N., S.K.; Resources- T.J.N., A.A.P., M.C.A., S.Z., E.E.E., A.K.; Data Curation- I.T.F., C.M.B., G.A.L.; Writing-Original Draft- D.J., I.T.F., G.A.L., F.M.J.J., S.R.S.; Writing-Review & Edit- D.J., I.T.F., G.A.L., F.M.J.J., S.R.S., C.M.B., A.D.E., G.L.M., A.R.F., M.H., T.J.N., A.A.P., M.L.D., X.N.; Visualization- I.T.F., G.A.L., G.L.M., A.M.N., M.H.; Supervision- A.K., E.E.E., S.R.S., F.M.J.J., D.H.; Project Administration- D.H., F.M.J.J.; Funding acquisition- D.H., F.M.J., E.E.E.

References

Bailey, J.A., and Eichler, E.E. (2006). Primate segmental duplications: crucibles of evolution, diversity and disease. Nat Rev Genet *7*, 552-564.

Bernier, R., Steinman, K.J., Reilly, B., Wallace, A.S., Sherr, E.H., Pojman, N., Mefford, H.C., Gerdts, J., Earl, R., Hanson, E.*, et al.* (2016). Clinical phenotype of the recurrent 1q21.1 copy-number variant. Genet Med *18*, 341-349.

Brunetti-Pierri, N., Berg, J.S., Scaglia, F., Belmont, J., Bacino, C.A., Sahoo, T., Lalani, S.R., Graham, B., Lee, B., Shinawi, M.*, et al.* (2008). Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities. Nat Genet *40*, 1466-1471.

Byrne, A., Beaudin, A.E., Olsen, H.E., Jain, M., Cole, C., Palmer, T., DuBois, R.M., Forsberg, E.C., Akeson, M., and Vollmers, C. (2017). Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. Nat Commun *8*, 16027.

Chang, H., Li, L., Peng, T., Li, M., Gao, L., and Xiao, X. (2016). Replication analyses of four chromosomal deletions with schizophrenia via independent large-scale meta-analyses. Am J Med Genet B Neuropsychiatr Genet *171*, 1161-1169.

Charrier, C., Joshi, K., Coutinho-Budd, J., Kim, J.E., Lambert, N., de Marchena, J., Jin, W.L., Vanderhaeghen, P., Ghosh, A., Sassa, T.*, et al.* (2012). Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation. Cell *149*, 923-935.

Cheng, W., Su, Y., and Xu, F. (2013). CHD1L: a novel oncogene. Mol Cancer *12*, 170.

Cheng, Z., Ventura, M., She, X., Khaitovich, P., Graves, T., Osoegawa, K., Church, D., DeJong, P., Wilson, R.K., Paabo, S.*, et al.* (2005). A genome-wide comparison of recent chimpanzee and human segmental duplications. Nature *437*, 88-93.

Dennis, M.Y., Nuttle, X., Sudmant, P.H., Antonacci, F., Graves, T.A., Nefedov, M., Rosenfeld, J.A., Sajjadian, S., Malig, M., Kotkiewicz, H.*, et al.* (2012). Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. Cell *149*, 912-922.

DeSilva, J., and Lesnik, J. (2006). Chimpanzee neonatal brain size: Implications for brain growth in Homo erectus. J Hum Evol *51*, 207-212.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21.

Doggett, N.A., Xie, G., Meincke, L.J., Sutherland, R.D., Mundt, M.O., Berbari, N.S., Davy, B.E., Robinson, M.L., Rudd, M.K., Weber, J.L.*, et al.* (2006). A 360-kb interchromosomal duplication of the human HYDIN locus. Genomics *88*, 762-771.

Dougherty, M.L., Nuttle, X., Penn, O., Nelson, B.J., Huddleston, J., Baker, C., Harshman, L., Duyzend, M.H., Ventura, M., Antonacci, F.*, et al.* (2017). The birth of a human-specific neural gene by incomplete duplication and gene fusion. Genome Biol *18*, 49.

Duan, Z., Li, F.Q., Wechsler, J., Meade-White, K., Williams, K., Benson, K.F., and Horwitz, M. (2004). A novel notch protein, N2N, targeted by neutrophil elastase and implicated in hereditary neutropenia. Mol Cell Biol *24*, 58-70.

Duering, M., Karpinska, A., Rosner, S., Hopfner, F., Zechmeister, M., Peters, N., Kremmer, E., Haffner, C., Giese, A., Dichgans, M.*, et al.* (2011). Co-aggregate formation of CADASIL-mutant NOTCH3: a single-particle analysis. Hum Mol Genet *20*, 3256-3265.

Dumas, L.J., O'Bleness, M.S., Davis, J.M., Dickens, C.M., Anderson, N., Keeney, J.G., Jackson, J., Sikela, M., Raznahan, A., Giedd, J.*, et al.* (2012). DUF1220-domain copy number implicated in human brain-size pathology and evolution. Am J Hum Genet *91*, 444-454.

Eiraku, M., and Sasai, Y. (2011). Mouse embryonic stem cell culture for generation of three-dimensional retinal and cortical tissues. Nat Protoc *7*, 69-79.

Eiraku, M., Watanabe, K., Matsuo-Takasaki, M., Kawada, M., Yonemura, S., Matsumura, M., Wataya, T., Nishiyama, A., Muguruma, K., and Sasai, Y. (2008). Self-organized formation of polarized cortical tissues from ESCs and its active manipulation by extrinsic signals. Cell Stem Cell *3*, 519-532.

Florio, M., Albert, M., Taverna, E., Namba, T., Brandl, H., Lewitus, E., Haffner, C., Sykes, A., Wong, F.K., Peters, J.*, et al.* (2015). Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. Science *347*, 1465-1470.

Girirajan, S., Dennis, M.Y., Baker, C., Malig, M., Coe, B.P., Campbell, C.D., Mark, K., Vu, T.H., Alkan, C., Cheng, Z.*, et al.* (2013). Refinement and discovery of new hotspots of copy-number variation associated with autism spectrum disorder. Am J Hum Genet *92*, 221-237.

Gollob, M.H., Jones, D.L., Krahn, A.D., Danis, L., Gong, X.Q., Shao, Q., Liu, X., Veinot, J.P., Tang, A.S., Stewart, A.F.*, et al.* (2006). Somatic mutations in the connexin 40 gene (GJA5) in atrial fibrillation. N Engl J Med *354*, 2677-2688.

Groot, A.J., Habets, R., Yahyanejad, S., Hodin, C.M., Reiss, K., Saftig, P., Theys, J., and Vooijs, M. (2014). Regulated proteolysis of NOTCH2 and NOTCH3 receptors by ADAM10 and presenilins. Mol Cell Biol *34*, 2822-2832.

Grozeva, D., Conrad, D.F., Barnes, C.P., Hurles, M., Owen, M.J., O'Donovan, M.C., Craddock, N., Kirov, G., and Wtccc (2012). Independent estimation of the frequency of rare CNVs in the UK population confirms their role in schizophrenia. Schizophr Res *135*, 1-7.

Habets, R.A., Groot, A.J., Yahyanejad, S., Tiyanont, K., Blacklow, S.C., and Vooijs, M. (2015). Human NOTCH2 Is Resistant to Ligand-independent Activation by Metalloprotease Adam17. J Biol Chem *290*, 14705-14716.

Haldeman-Englert, C.R., and Jewett, T. (1993). 1q21.1 Recurrent Microdeletion. In GeneReviews(R), M.P. Adam, H.H. Ardinger, R.A. Pagon, S.E. Wallace, L.J.H. Bean, H.C. Mefford, K. Stephens, A. Amemiya, and N. Ledbetter, eds. (Seattle (WA)).

Hansen, D.V., Lui, J.H., Parker, P.R., and Kriegstein, A.R. (2010). Neurogenic radial glia in the outer subventricular zone of human neocortex. Nature *464*, 554-561.

Holloway, R., Broadfield, D., and Yuan, M. (2004). THE HUMAN FOSSIL RECORD, Volume Three: Brain Endocasts-The Paleoneurological Evidence, Vol 3 (Hoboken, NJ: John Wiley & Sons).

International Schizophrenia Consortium (2008). Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature *455*, 237-241.

Jain, M., Fiddes, I.T., Miga, K.H., Olsen, H.E., Paten, B., and Akeson, M. (2015). Improved data analysis for the MinION nanopore sequencer. Nat Methods *12*, 351-356.

Joutel, A., Vahedi, K., Corpechot, C., Troesch, A., Chabriat, H., Vayssiere, C., Cruaud, C., Maciazek, J., Weissenbach, J., Bousser, M.G.*, et al.* (1997). Strong clustering and stereotyped nature of Notch3 mutations in CADASIL patients. Lancet *350*, 1511-1515.

Ju, X.C., Hou, Q.Q., Sheng, A.L., Wu, K.Y., Zhou, Y., Jin, Y., Wen, T., Yang, Z., Wang, X., and Luo, Z.G. (2016). The hominoid-specific gene TBC1D3 promotes generation of basal neural progenitors and induces cortical folding in mice. Elife *5*.

Kageyama, R., Ohtsuka, T., Shimojo, H., and Imayoshi, I. (2009). Dynamic regulation of Notch signaling in neural progenitor cells. Curr Opin Cell Biol *21*, 733-740.

Karlstrom, H., Beatus, P., Dannaeus, K., Chapman, G., Lendahl, U., and Lundkvist, J. (2002). A CADASIL-mutated Notch 3 receptor exhibits impaired intracellular trafficking and maturation but normal ligand-induced signaling. Proc Natl Acad Sci U S A *99*, 17119-17124.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol *14*, R36.

Kopan, R., and Ilagan, M.X. (2009). The canonical Notch signaling pathway: unfolding the activation mechanism. Cell *137*, 216-233.

Lancaster, M.A., Renner, M., Martin, C.A., Wenzel, D., Bicknell, L.S., Hurles, M.E., Homfray, T., Penninger, J.M., Jackson, A.P., and Knoblich, J.A. (2013). Cerebral organoids model human brain development and microcephaly. Nature *501*, 373-379.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357-359.

Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P.H., Schraiber, J.G., Castellano, S., Lipson, M.*, et al.* (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature *513*, 409-413.

Lewitus, E., Kelava, I., and Huttner, W.B. (2013). Conical expansion of the outer subventricular zone and the role of neocortical folding in evolution and development. Front Hum Neurosci *7*, 424.

Louvi, A., and Artavanis-Tsakonas, S. (2006). Notch signalling in vertebrate neural development. Nat Rev Neurosci *7*, 93-102.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol *15*, 550.

Lui, J.H., Hansen, D.V., and Kriegstein, A.R. (2011). Development and evolution of the human neocortex. Cell *146*, 18-36.

Mace, A., Tuke, M.A., Deelen, P., Kristiansson, K., Mattsson, H., Noukas, M., Sapkota, Y., Schick, U., Porcu, E., Rueger, S.*, et al.* (2017). CNV-association meta-analysis in 191,161 European adults reveals new loci associated with anthropometric traits. Nat Commun *8*, 744.

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A.*, et al.* (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature *538*, 201-206.

Marchetto, M.C.N., Narvaiza, I., Denli, A.M., Benner, C., Lazzarini, T.A., Nathanson, J.L., Paquola, A.C.M., Desai, K.N., Herai, R.H., Weitzman, M.D.*, et al.* (2013). Differential L1 regulation in pluripotent stem cells of humans and apes. Nature *503*, 525-529.

Mefford, H.C., Sharp, A.J., Baker, C., Itsara, A., Jiang, Z., Buysse, K., Huang, S., Maloney, V.K., Crolla, J.A., Baralle, D.*, et al.* (2008). Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. N Engl J Med *359*, 1685-1699.

Meyer, M., Kircher, M., Gansauge, M.T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prufer, K., de Filippo, C.*, et al.* (2012). A high-coverage genome sequence from an archaic Denisovan individual. Science *338*, 222-226.

Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., and Thomas, P.D. (2017). PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Res *45*, D183-D189.

Molnar, Z., Metin, C., Stoykova, A., Tarabykin, V., Price, D.J., Francis, F., Meyer, G., Dehay, C., and Kennedy, H. (2006). Comparative aspects of cerebral cortical development. Eur J Neurosci *23*, 921-934.

Nichols, J.T., Miyamoto, A., Olsen, S.L., D'Souza, B., Yao, C., and Weinmaster, G. (2007). DSL ligand endocytosis physically dissociates Notch1 heterodimers before activating proteolysis can occur. J Cell Biol *176*, 445-458.

Nuttle, X., Giannuzzi, G., Duyzend, M.H., Schraiber, J.G., Narvaiza, I., Sudmant, P.H., Penn, O., Chiatante, G., Malig, M., Huddleston, J.*, et al.* (2016). Emergence of a Homo sapiens-specific gene family and chromosome 16p11.2 CNV susceptibility. Nature *536*, 205-209.

Nuttle, X., Huddleston, J., O'Roak, B.J., Antonacci, F., Fichera, M., Romano, C., Shendure, J., and Eichler, E.E. (2013). Rapid and accurate large-scale genotyping of duplicated genes and discovery of interlocus gene conversions. Nat Methods *10*, 903-909.

O'Bleness, M., Searles, V.B., Dickens, C.M., Astling, D., Albracht, D., Mak, A.C., Lai, Y.Y., Lin, C., Chu, C., Graves, T.*, et al.* (2014). Finished sequence and assembly of the DUF1220-rich 1q21 region using a haploid human genome. BMC Genomics *15*, 387.

O'Bleness, M., Searles, V.B., Varki, A., Gagneux, P., and Sikela, J.M. (2012). Evolution of genetic and genomic features unique to the human lineage. Nat Rev Genet *13*, 853-866.

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.

Pasca, A.M., Sloan, S.A., Clarke, L.E., Tian, Y., Makinson, C.D., Huber, N., Kim, C.H., Park, J.Y., O'Rourke, N.A., Nguyen, K.D.*, et al.* (2015). Functional cortical neurons and astrocytes from human pluripotent stem cells in 3D culture. Nat Methods *12*, 671-678.

Peng, Y., Leung, H.C., Yiu, S.M., and Chin, F.Y. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics *28*, 1420-1428.

Pollen, A.A., Nowakowski, T.J., Chen, J., Retallack, H., Sandoval-Espinosa, C., Nicholas, C.R., Shuga, J., Liu, S.J., Oldham, M.C., Diaz, A.*, et al.* (2015). Molecular identity of human outer radial glia during cortical development. Cell *163*, 55-67.

Ponce de Leon, M.S., Golovanova, L., Doronichev, V., Romanova, G., Akazawa, T., Kondo, O., Ishida, H., and Zollikofer, C.P. (2008). Neanderthal brain size at birth provides insights into the evolution of human life history. Proc Natl Acad Sci U S A *105*, 13764-13768.

Popesco, M.C., Maclaren, E.J., Hopkins, J., Dumas, L., Cox, M., Meltesen, L., McGavran, L., Wyckoff, G.J., and Sikela, J.M. (2006). Human lineage-specific amplification, selection, and neuronal expression of DUF1220 domains. Science *313*, 1304-1307.

Prufer, K., de Filippo, C., Grote, S., Mafessoni, F., Korlevic, P., Hajdinjak, M., Vernot, B., Skov, L., Hsieh, P., Peyregne, S.*, et al.* (2017). A high-coverage Neandertal genome from Vindija Cave in Croatia. Science *358*, 655-658.

Prufer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P.H., de Filippo, C.*, et al.* (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. Nature *505*, 43-49.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Rabouille, C. (2017). Pathways of Unconventional Protein Secretion. Trends Cell Biol *27*, 230-240.

Rakic, P. (2009). Evolution of the neocortex: a perspective from developmental biology. Nat Rev Neurosci *10*, 724-735.

Rosenfeld, J.A., Traylor, R.N., Schaefer, G.B., McPherson, E.W., Ballif, B.C., Klopocki, E., Mundlos, S., Shaffer, L.G., Aylsworth, A.S., and q21.1 Study, G. (2012). Proximal microdeletions and microduplications of 1q21.1 contribute to variable abnormal phenotypes. Eur J Hum Genet *20*, 754-761.

Shiels, A., Mackay, D., Ionides, A., Berry, V., Moore, A., and Bhattacharya, S. (1998). A missense mutation in the human connexin50 gene (GJA8) underlies autosomal dominant "zonular pulverulent" cataract, on chromosome 1q. Am J Hum Genet *62*, 526-532.

Skarnes, W.C. (2000). Gene trapping methods for the identification and functional analysis of cell surface proteins in mice. Methods Enzymol *328*, 592-615.

Smart, I.H., Dehay, C., Giroud, P., Berland, M., and Kennedy, H. (2002). Unique morphological features of the proliferative zones and postmitotic compartments of the neural epithelium giving rise to striate and extrastriate cortex in the monkey. Cereb Cortex *12*, 37-53.

Stankiewicz, P., and Lupski, J.R. (2010). Structural variation in the human genome and its role in disease. Annu Rev Med *61*, 437-455.

Steinberg, K.M., Schneider, V.A., Graves-Lindsay, T.A., Fulton, R.S., Agarwala, R., Huddleston, J., Shiryev, S.A., Morgulis, A., Surti, U., Warren, W.C.*, et al.* (2014). Single haplotype assembly of the human genome from a hydatidiform mole. Genome Res *24*, 2066-2076.

Steringer, J.P., Muller, H.M., and Nickel, W. (2015). Unconventional secretion of fibroblast growth factor 2--a novel type of protein translocation across membranes? J Mol Biol *427*, 1202-1210.

Sudmant, P.H., Kitzman, J.O., Antonacci, F., Alkan, C., Malig, M., Tsalenko, A., Sampas, N., Bruhn, L., Shendure, J., Genomes, P.*, et al.* (2010). Diversity of human copy number variation and multicopy genes. Science *330*, 641-646.

Szamalek, J.M., Goidts, V., Cooper, D.N., Hameister, H., and Kehrer-Sawatzki, H. (2006). Characterization of the human lineage-specific pericentric inversion that distinguishes human chromosome 1 from the homologous chromosomes of the great apes. Hum Genet *120*, 126-138.

Van Dijck, A., van der Werf, I.M., Reyniers, E., Scheers, S., Azage, M., Siefkas, K., Van der Aa, N., Lacroix, A., Rosenfeld, J., Argiropoulos, B.*, et al.* (2015). Five patients with a chromosome 1q21.1 triplication show macrocephaly, increased weight and facial similarities. Eur J Med Genet *58*, 503-508.

Varki, A., Geschwind, D.H., and Eichler, E.E. (2008). Explaining human uniqueness: genome interactions with environment, behaviour and culture. Nat Rev Genet *9*, 749-763.

Xu, C., Aragam, N., Li, X., Villla, E.C., Wang, L., Briones, D., Petty, L., Posada, Y., Arana, T.B., Cruz, G.*, et al.* (2013). BCL9 and C9orf5 are associated with negative symptoms in schizophrenia: meta-analysis of two genome-wide association studies. PLoS One *8*, e51674.

Xu, X., Choi, S.H., Hu, T., Tiyanont, K., Habets, R., Groot, A.J., Vooijs, M., Aster, J.C., Chopra, R., Fryer, C.*, et al.* (2015). Insights into Autoregulation of Notch3 from Structural and Functional Studies of Its Negative Regulatory Region. Structure *23*, 1227-1235.

Zerbino, D.R., Johnson, N., Juettemann, T., Wilder, S.P., and Flicek, P. (2014). WiggleTools: parallel processing of large collections of genome-wide datasets for visualization and statistical analysis. Bioinformatics *30*, 1008-1009.

Supplemental Figure Legends (See "Supplemental Materials.pdf" for the figures)

Figure S1. Related to Figure 1. (A) Identity between *NOTCH2*-related genes as measured over the alignable genomic region (blue) or the gene exons (pink). (B) DNA and amino acid sequence of human NOTCH2NL genes in exon 5, which is derived from *NOTCH2* intron 4. (C) Immunoblot of mouse ESCs transfected with WT *NOTCH2NL*, or *NOTCH2NL* with ancestral ATAA inserted in exon 5. (D) Relative protein levels based on the immunoblot in panel C. (E) RT-qPCR analysis of the same samples for determination of transcript levels for each condition. (F) Alignment of the H9 assembled paratypes to GRCh38. Each paratype is colored as to whether a position aligns best to GRCH38 *NOTCH2NLA*, *NOTCH2NLB* or *NOTCH2NLC*. (G) Observed frequency of individuals with the indicated *NOTCH2NLC* and *NOTCH2NLR* copy number in the Simons Diversity Panel (n=266). (H) Schematic of linked-read sequencing and Gordian Assembler protocol using the 10X genomics Chromium genome assay and oligo-capture to enrich for library fragments containing the desired genomic region. (I) Protein alignment of observed NOTCH2, NOTCH2NL and NOTCH2NLR paratypes based on our assembly results. Note that NOTCH2 sequence extends beyond what is shown in the alignment. A segregating variant in NOTCH2NLR is found at amino acid position 235.

Figure S2. Related to Figure 2. Details of *NOTCH2NL*-like genes in Gorilla (A) and Chimp (B) and source of genome sequence support. (C) Verification of fusion genes by RT-PCR on Chimp and Gorilla RNA. H = Human, C = Chimpanzee, G = Gorilla. (D) Immunoblot blot using an N-terminal NOTCH2 antibody (aa 25-255), comparing ectopic expression of Chimp *NOTCH2NL*-like gene lacking exon 2 compared to ectopic expression of human *NOTCH2NL*$^{Sh,\ T197I}$ in mESCs. (E) Endocranial volume of fossil

hominids versus time as determined by Holloway, et al. 2004. (F) Details of splice junctions of fusion genes and related open reading frames. Top rows show nucleotide sequence in fusion transcripts. Middle rows show peptides derived from these transcripts. Lower rows contain the orthologous human NOTCH2NL protein sequence.

Figure S3. Related to Figure 3. (A) Summary violin plots indicating *NOTCH2NL* and *NOTCH2* expression in various cell types. (B) *NOTCH2NL* paratype expression in undifferentiated hESCs and week 5 cortical organoids from bulk Illumina RNA-Seq.

Figure S4. Related to Figure 4. (A) MA plot of RNA-sequencing data of mouse cortical organoids based on DESeq2 analysis. (B) Z-score of differentially expressed genes (p-adj < 0.05, DESeq2). (C) GO terms significantly associated with the upregulated genes in organoids ectopically expressing *NOTCH2NL*^{Sh,T197I}.

Figure S5. Related to Figure 5. (A) Schematic of strategy to generate *NOTCH2NL*-specific deletions using CRISPR/Cas9 and alignment of the two guide sequences used to NOTCH2NL-related sequences. (B) Heatmap of expression levels for a selection of brain structure marker genes from hESC-derived cortical organoids at the indicated time points (Left) and from human embryonic dorsal prefrontal cortex (DFC) samples at 8 pcw, 9 pcw, 12 pcw and 13 pcw, derived from the Allen Brain Atlas (http://www.brainspan.org). w = week; pcw = post conception week; FPKM= fragments per kilobase of exon per million fragments mapped. (C) Brightfield images of developing H9* and H9$^{NOTCH2NL\Delta}$ cortical organoids. Day 28 images show organoid pools for 1, RNA-Seq replicate.

Figure S6. Related to Figure 6. (A-B) Investigation of co-immunoprecipitation of NOTCH2NL with PDGFRB and EGFR in two independent experiments. (C) Reporter assay to assess the effect of NOTCH2NL using either NOTCH1-GAL4, NOTCH2-GAL4 or NOTCH3-GAL4 to induce pGL3-UAS reporter activation. 6 replicates in one experiment. Student's t-test with Holm-Bonferroni correction (* $p < 0.05$, ** $p < 10^{-3}$, *** $p < 10^{-5}$), error bars indicate SD. (D) NOTCH2NL effects of NOTCH2-GAL4 reporter assay co-culture with JAG2 or DLL1 expressing cells. 6 replicates in one experiment. Student's t-test with Holm-Bonferroni correction (* $p<0.05$, ** $p<10^{-3}$, *** $p<10^{-5}$), error bars indicate SD. (E) NOTCH2NL remains effective in an assay with activation of NOTCH2-GAL4 using recombinant DLL4 coated plates. Average of 3 independent experiments with 4 or 5 replicates each. Two-way anova with Tukey's HSD (* $p < 10^{-4}$, ** $p <10^{-8}$, *** $p < 10^{-12}$), error bars indicate SEM.

Figure S7. Related to Figure 7. Relative probe intensities from CNV-microarrays for 2 controls and 11 patients with reported 1q21.1 aberrations mapped to the GRCh38 1q21.1 assembly. Gray: normal, red: deletion, blue: duplication. Dark red/blue is high confidence deletion/duplication based on probe values, light red/blue are potentially part of the deletion/duplication.

Table S1. Related to Figures 1. Results of NOTCH2NL gene de novo assembly.

Table S2. Related to Figure 4 and 5. Gene expression measurements from RNA-seq experiments.

Table S3. Related to Figure 7. Features of Simons VIP samples.

Table S4. Related to Figures 1 and 7. Curated Paratypes of Assembled and Simons Normals

STAR Methods

Mouse stem cell culture and organoid differentiation

Mouse 46c ESCs were grown according to the BayGenomics protocol (Skarnes, 2000), on 0.1% gelatin coated plates, and cultured in GMEM (Thermofisher) with 10% HIFBS, 2 mM L-Glutamine, 1x NEAA, 1x NaPyr, 100 µM 2-mercaptoethanol and 1x P/S. ESGRO LIF (Millipore) was added fresh daily. To generate stable cell lines, 46C cells seeded on 100 mm plates and were transfected with 24 µg of linearized pCIG-NOTCH2NL-ires-GFP or empty pCIG-ires-GFP vector, using lipofectamine 2000 (Thermofisher). After 36 hours, GFP-positive cells were sorted using a FACSAria III (BD Biosciences) and recovered for further culturing. After 4 passages sorting was repeated and GFP-positive cells that had stably integrated the plasmid DNA in their genome were recovered for expansion and further culturing. We verified continued stable expression of NOTCH2NL-ires-GFP or empty vector (Supplemental Figure S4). Mouse 46C ESC organoid differentiation was performed as described previously (Eiraku and Sasai, 2011). Three pools of 16 organoids were isolated in TRIzol after 6 days of differentiation for EV and NOTCH2NL organoids.

Human stem cell culture

Mitomycin-C treated mouse embryonic fibroblasts (MEFs, GlobalStem) were seeded on 0.1% gelatin coated plates at a density of 35.000 cell/cm^2. MEFs were cultured in DMEM, 4.5 g/l glucose + GlutaMax (Thermofisher, 10% heat inactivated fetal bovine serum (HIFBS, Thermofisher), 1x Penicillin/Streptomycin (P/S, Thermofisher) and 1x sodium pyruvate (NaPyr, Thermofisher). H9 human embryonic stem cells (H9 hESCs, WiCell), were cultured in W0 medium: DMEM/F12 (Thermofisher) with 20% KnockOut serum replacement (KOSR, Thermofisher), 2 mM L-glutamine (Thermofisher), 1x non-essential amino acids (NEAA, Thermofisher), 100 uM 2-mercaptoethanol (Thermofisher) and 1x P/S (Thermofisher). W0 was

freshly supplemented daily with 8 ng/ml FGF2 (Sigma). H9 hESCs were grown on MEF feeder layers, and manually passaged every 5-6 days when colonies reached approximately 2 mm in diameter.

Human organoid differentiation

For organoid differentiation, medium was replaced with W0 medium + 1x NaPyr without FGF2 (Differentiation medium). Colonies of 2-3 mm in diameter were manually lifted using a cell lifter, and transferred to an ultra-low attachment 60mm dish (Corning). After 24 hours (day 0) embryoid bodies had formed, and 50% of medium was replaced with Differentiation medium supplemented with small molecule inhibitors and recombinant proteins to the following final concentrations: 500 ng/ml DKK1 (peprotech), 500 ng/ml NOGGIN (R&D Systems), 10 µM SB431542 (Sigma) and 1 µM Cyclopamine V. californicum (VWR). Medium was then replaced every other day until harvest. On day 8, organoids were transferred to ultra-low attachment U-shaped bottom 96 well plates (Corning). On day 18, medium was changed to Neurobasal/N2 medium: Neurobasal (Thermofisher), 1x N2 supplement (Thermofisher), 2 mM L-Glutamine, 1x P/S, supplemented with 1 µM Cyclopamine. From day 26 on, Cyclopamine was not supplemented anymore. Organoids were harvested in TRIzol at weekly time points. Total-transcriptome strand-specific RNA sequencing libraries were generated using dUTP for second strand synthesis on Ribo-zero depleted total RNA (Parkhomchuk et al., 2009). For organoid formation of H9 hESC CRISPR/Cas9 NOTCH2NL knockout lines, an updated protocol was used: Differentiation medium was supplemented with 10 µM SB431542 (Sigma), 1 µM Dorsomorphin (Sigma), 3 µM IWR-1-Endo (Sigma) and 1 µM Cyclopamine (Sigma). Medium was then replaced every other day until harvest. On day 4, 60 mm dishes with organoids were placed on a hi/lo rocker in the incubator. From day 18 on, medium is replaced with Neurobasal/N2 medium. From day 24 on, Cyclopamine was not added anymore. Pools of 5-10 organoids per replicate were harvested in TRIzol at day 28 for RNA extraction.

RNA-Sequencing Analysis

Paired-end Illumina reads were trimmed from the 3' end of read1 and read2 to 100x100 bp for human. Bowtie2 v2.2.1 (Langmead and Salzberg, 2012) was used with the "--very-sensitive" parameter to filter reads against the RepeatMasker library (http://www.repeatmasker.org) which were removed from further analysis. STAR v2.5.1b (Dobin et al., 2013) was used to map RNA-seq reads to the human reference

genome GRCh37. STAR was run with the default parameters with the following exceptions: --outFilterMismatchNmax 999, --outFilterMismatchNoverLmax 0.04, --alignIntronMin 20, --alignIntronMax 1000000, and --alignMatesGapMax 1000000. STAR alignments were converted to genomic position coverage with the bedtools command genomeCoverageBed –split (Quinlan and Hall, 2010). DESeq2 v1.14.1 (Love et al., 2014)) was used to provide basemean expression values and differential expression analysis across the time course. Total gene coverage for a gene was converted to read counts by dividing the coverage by N+N (100+100) since each paired-end NxN mapped read induces a total coverage of N+N across its genomic positions. Results are in **Table S2** and data are available from GEO: GSE106245.

For mouse cortical organoids, H9* and H9$^{NOTCH2NL\Delta}$ organoid samples, RNA was isolated according to standard TRIzol protocol. RNA was treated with DNAseI (Roche) according to standard protocol for DNA clean-up in RNA samples. RNA was then isolated by column purification (Zymo RNA clean & concentrator 5) and stored at -80°C. For RNA sequencing, mRNA was isolated from total RNA using polyA selection Dynabeads mRNA DIRECT Micro Purification Kit (Thermofisher). Library was prepared using strand-specific Ion Total RNA-seq Kit v2 (Thermofisher) and Ion Xpress RNA-seq Barcode 1-16 (Thermofisher) to label different samples. The samples were sequenced on an IonProton system (Thermofisher), generating single-end reads of around 100 bp in length. IonProton RNA sequencing data was processed using the Tuxedo package, in consideration of parameters recommended by Thermofisher for IonProton data. Briefly, samples were mapped using Tophat2 (Kim et al., 2013), using Bowtie2 (Langmead and Salzberg, 2012) as the underlying alignment tool. The target genome assembly for these samples was GRCh38/UCSC hg38, and Tophat was additionally supplied with the gene annotation of ENSEMBL84 (GRCh38.p5). Reads mapped per exon were counted using HT-Seq count (union) and summed per corresponding gene. HT-Seq count output was normalized using DESeq2, and pairwise comparisons were made to determine significant differences in gene expression. Results are in **Table S2**.

For comparison of week 4 H9* and H9$^{NOTCH2NL\Delta}$ organoid data to the previously established H9 organoid timeline, the following procedure was used: The top 250 upregulated and the 250 downregulated genes

between week 4 H9* and H9$^{NOTCH2NL\Delta}$ based on p-adj were selected. The matching expression profiles of these 500 genes were extracted from the H9 organoid timeline, yielding 361 genes expressed in both datasets. The expression profiles in week 4 H9* and H9$^{NOTCH2NL\Delta}$ and H9 Week 3, Week 4 and Week 5 were sorted from high to low, and ranked 1 to 361. Then, pairwise comparisons were made between each sample to calculate Spearman's rank correlation between all samples, and plotted using multi-experiment viewer. 212 genes showed shift towards better correlation with Week 5 data in H9$^{NOTCH2NL\Delta}$ compared to H9*. These 212 genes were subjected to GO analysis using Panther V13.0 (Mi et al., 2017). A selection of genes from the significantly associated term neuron differentiation was plotted in a heatmap. Z-scores were calculated for the different samples of Week 4 H9* and H9$^{NOTCH2NL\Delta}$, and H9 Week 3, Week 4 and Week 5.

Co-immunoprecipitation and immunoblot

HEK293T cells were cultured according to standard protocol in DMEM, 4.5 g/l glucose + GlutaMax, 10% HIFBS and 1x P/S. For co-IP experiments, HEK293T cells were transfected in T25 flasks at 50% confluency. pCIG-NOTCH2-Myc and pCAG-NOTCH2NL-HASh + pCAG-NOTCH2NL-HAL,T197I were mixed in equimolar ratios and transfected using Lipofectamine 2000 (Thermofisher). For control conditions, pCIG-EV and pCAG-EV were used in equimolar ratios. 6 hours after transfection, medium was replaced, and another 24 hours later medium was replaced. Cells were harvested 48 hours after transfection. Cells were washed 3 times with cold 1x PBS, then incubated in 40 minutes in IP buffer (50mM Tris-HCl, 150mM NaCl, 5mM MgCl, 0.5mM EDTA, 0.2% NP-40, 5% glycerol, supplemented with cOmplete, EDTA-free protease inhibitor cocktail (Sigma)) Cells were lysed by passing cell suspension through 273/4 gauge needle 10 times. Lysate was centrifuged 10 minutes at 4°C, supernatant was transferred to a fresh 1.5ml tube. 2 µg of one specific antibody was added (anti-HA Abcam ab9110, anti-Myc Abcam ab9E10, anti-His Abcam ab9108, anti-NOTCH2 SCBT sc25-255) and incubated overnight at 4°C in a rotating wheel. DynaBeads were blocked using 3 washed of 1x PBS + 0.5% BSA and added to IP samples, incubating 3 hours at 4°C rotating. Using a magnetic separator, samples were washed 2 times in cold IP buffer. Then samples were eluted in Tris-EDTA buffer and transferred to new 1.5ml tubes. 2x Laemmli buffer + DTT was added 1:1 prior to SDS PAGE. Samples were loaded on 4-20% Tris glycine

gels (Bio-Rad), followed by blotting on nitrocellulose membranes following manufacturer's recommended protocol. Membranes were blocked in 5% skim-milk powder in 1x PBS + 0.05% Tween or 1x TBS + 0.1% Tween. Primary antibodies were incubated 3 hours at room temperature in 1x PBS (anti-NOTCH2 sc25-255) or 1x TBS-T (other antibodies), followed by 3 washes in 1x PBS-T (anti-NOTCH2 sc25-255) or 1x TBS-T (other antibodies). Secondary antibodies (anti-Rabbit-HRP, anti-Mouse-HRP, Thermofisher) were incubated 60 minutes at room temperature, followed by 3 more washes in 1x PBS-T or 1x TBS-T. Membranes were incubated with supersignal westdura ECL substrate (Thermofisher) and imaged using Bio-Rad Chemidoc imager. For experiments with pCAG-NOTCH2NL-His, pCIG-NOTCH2-Myc, pCIG-PDGFRB-Myc and pCIG-EGFR-Myc, the same protocol was used with equimolar mixes of plasmid DNA. For immunoprecipitation of NOTCH2NLSh,T197I from mouse 46c ESCs, the same protocol was used and protein was isolated from medium using the NOTCH2 sc25-255 antibody.

NOTCH reporter cell line culture

U2OS cells were cultured in DMEM, 4.5 g/l glucose + GlutaMax, 10% HIFBS and 1x P/S. U2OS-JAG2 cells (gift of Arjan Groot and Marc Vooijs, MAASTRO lab, Maastricht University) were supplemented with 2 µg/ml puromycin. OP9 cells and OP9-DLL1 (gift of Bianca Blom, Academic Medical Center Amsterdam) were cultured in MEMα without nucleosides (Thermofisher), 2mM L-glutamine, 20% HIFBS, 100 µM 2-mercaptoethanol and 1x P/S. For routine culturing, cells were passaged every 3-4 days using 0.25% Trypsin (Thermofisher) + 0.5 mM EDTA (Sigma) in PBS at densities of 1/8 to 1/10 (U2OS), or 1/4 to 1/6 (OP9).

NOTCH reporter co-culture assays

U2OS cells were seeded at a density of 425,000 cells per well for transfection (6-wells plate). In parallel, U2OS control or U2OS-JAG2 cells were seeded at a density of 110,000 cells per well for co-culture (12-wells plate). After 24 hours, U2OS cells in 6-wells plates were transfected the following amounts of plasmid DNA per well. For control conditions: ng pGL3-UAS, 33.3 ng CMV-Renilla, 16.7 ng pCAG-GFP, 200 ng pcDNA5.1-NOTCH2-GAL4, 167 ng pCAG-EV, and 273 ng pBluescript. For conditions including NOTCH2NL:500 ng pGL3-UAS, 33.3 ng CMV-Renilla, 16.7 ng pCAG-GFP, 200 ng pcDNA5.1-NOTCH2-GAL4, 200 ng pCAG-NOTCH2NL, and 240 ng pBluescript. Plasmid DNA mix was transfected using

polyethylenimine (PEI, linear, MW 25000, Polysciences). All amounts were scaled accordingly for multiple transfections. For larger experiments, cells were seeded and transfected in T25 flasks or on 100 mm plates and amounts used were scaled accordingly to surface area. 6 hours after transfection, 6-wells plates were treated with 0.5 ml of 0.25% Trypsin and 0.5 mM EDTA in PBS per well for 2 minutes at 37 degrees. Cells were resuspended in a total volume of 7 ml after addition of culture medium. Medium of 12-wells plates was removed, and 1 ml of transfected cell suspension was added to each well for co-culture. 10 μM DBZ was added to selected control wells. After 24 hours, medium was removed and cells washed once with PBS. Cells were incubated in 150 μl of 1x passive lysis buffer (PLB, Promega) on an orbital shaker for 15 minutes. Lysates were stored at -80°C until analysis. In OP9 and OP9-DLL1 co-cultures, 80,000 cells were seeded per well of a 12-wells plate. For generating conditioned medium, U2OS cells were seeded on 100 mm plates, and were PEI transfected with 2000 ng of pCIG-EV, or 2400 ng of NOTCH2NLA or NOTCH2NLB. Another 10000 ng or 9600 ng of pBluescript was used as carrier DNA. 6 hours after transfection, medium was replaced. 32 hours after transfection, medium was collected and 0.2 μm filtered and used the same day. The experiments using conditioned medium were done as previously described, but were seeded on 0.25% gelatin, 0.1% BSA coated plates instead. For the reporter U2OS cell transfection, only pCAG-EV, and NOTCH2NL plasmids were not added to the plasmid DNA mix, and replaced by pBluescript. Instead, transfected cells are resuspended and seeded in conditioned medium harvested from other cells. For DLL4 assays, 24-wells plates were coated overnight at 4°C with 150 μl of 5 μg / ml rDLL4 (R&D Systems), 0.25% gelatin, 0.1% BSA in PBS. Control plates were coated with 0.25% gelatin, 0.1% BSA in PBS only. U2OS cells were transfected and seeded according to co-culture protocol as previously described, except 0.5 ml of cell suspension was used for each well of the coated 24-wells plates. NOTCH-GAL4 and reporter constructs were kindly gifted by Arjan Groot and Marc Vooijs (MAASTRO lab, Maastricht University).

RT-PCR characterization of primate *NOTCH2NL* fusion genes

For amplification and detection of potential fusion transcripts, Qiagen one-step RT-PCR kit was used according to manufacturer's protocol. 25 ng of total RNA isolated from gorilla iPSCs, chimpanzee iPSCs, or human H9 ESCs was used per reaction. Primers used in these reactions were:

Human_N2NL_Fw1_exon1: CGCTGGGCTTCGGAGCGTAG

Human_N2NL_Fw2_exon2: AGTGTCGAGATGGCTATGAA

Human_N2NL_Fw3_exon3: ATCGAGACCCCTGTGAGAAGA

Human_N2NL_Rv2_exon5: CCAGTGTCTAATTCTCATCG

PDE4DIP_Fw1_exon27: AAGGCCCAGCTGCAGAATGC

PDE4DIP_Fw2_exon24: ACACCATGCTGAGCCTTTGC

Chimp_N2NL_Rv1_exon3: GCAAGGTCGAGACACAGAGC

MAGI3_Fw1_exon1: GGGTTCGGGATGTCGAAGAC

MAGI3_Fw2_exon10: GCAACTGTGTCCTCGGTCAC

MAGI3_Fw3_exon14: GGGAGCAGCTGAGAAAGATG

TXNIP_Fw1_exon1: CAGTTCATCATGGTGATG

TXNIP_Fw2_exon1: GGGTACTTCAATACCTTGCATG

BRD9_Fw1_exon12: GCAGGAGTTTGTGAAGGATGC

BRD9_Fw2_exon10: ACGCTGGGCTTCAAAGACG

10x Library Enrichment

To enrich whole-genome sequencing libraries to allow for cost-effective deep sequencing of the *NOTCH2NL* loci, a MYcroarray MyBaits custom oligonucleotide library was developed. 100 bp probes were designed spaced 50 bp apart in chr1:145,750,000-149,950,000, ignoring repeat masked bases, for a total of 20,684 probes. A further 8,728 probes were created in the three *NOTCH2NL* loci by tiling with 50 bp overlaps, ignoring repeat masking but dropping any probes with very low complexity. 17,866 probes were added at every SUN position tiling at 5 bp intervals from -75 bp to +75 bp around the SUN. To try and capture population diversity and ensure even enrichment, at every SNP in the NA12878

Genome In a Bottle variant call set the reference base was replaced and probes tiled in the same fashion as the SUNs. Finally, to reach the required 60,060 probes a random 347 probes were dropped.

Library Preparation and Enrichment of 10x Chromium Libraries

High molecular weight DNA was processed into Illumina sequencing libraries using the Chromium Genome Reagent Kit V2 chemistry and enriched using the custom MyBaits oligonucleotide probes described above (**Figure S1**). Briefly, high molecular weight (HMW) gDNA was isolated from cultured cells using a MagAttract kit (Qiagen) followed by quantification with Qubit. HMW DNA was partitioned inside of an emulsion droplet along with DNA barcode containing gel beads and an amplification reaction mixture. After barcoding the molecules within the emulsion, Illumina sequencing adaptors were added by ligation. In preparation for hybridization with MyBaits probes Illumina adaptor sequences are blocked with complementary oligonucleotides. Biotinylated probes were hybridized overnight at 65C and isolated using streptavidin coated MyOne C1 beads (Invitrogen). The final enriched libraries were amplified using an Illumina Library Amplification Kit (Kapa).

Sequencing of Enriched 10x Libraries

The MYcroarray probes were used to enriched 10x Genomics sequencing libraries for three well studied individuals (NA19240, NA12877 and CHM1), the H9 ESC line, the six Simons VIP samples in Figure 7, and the H9 CRISPR mutant in Figure 5. NA12877 was chosen instead of NA12878 because of the existence of high depth 10x whole-genome data for that individual (see below). We find that around 50% of our reads map to regions of enrichment, leading to >1000x coverage of the *NOTCH2NL* loci. The NA19240, NA12877 and H9 libraries were sequenced to 65 million reads, 71 million reads, and 107 million reads respectively. The Simons VIP samples SV721, SV877, SV7720, SV780, SV735 and SV788 were sequenced to a depth of 57 million, 30 million, 44 million, 37 million, 86 million and 37 million reads respectively.

NOTCH2NL Simons Samples Coverage Analysis

To assess copy number change in the Simons VIP 1q21.1 collection, the H9, NA12877 and NA19240 normal libraries described above were mapped to GRCh38 using Longranger 2.1.3. bamCoverage was

used to extract all reads that mapped to the region chr1:142785299-150598866, normalizing depth to 1x coverage across the region to account for library depth. Wiggletools mean (Zerbino et al., 2014) was used to average the depth across these samples. Wiggletools was then used to perform a ratio of this average with the coverages of every Simons 1q21.1 collection sample, which simultaneously normalizes out bias from the array enrichment as well as GC content. These coverages were then re-scaled by the average coverage in the region chr1:149,578,286-149,829,369, which is downstream of *NOTCH2NLC* and not observed to have copy number change. This rescaling adjusts for a systematic shift downward caused by the combination of the previous normalizations seen in deletion samples, and a similar shift upward in duplication samples. Finally, sliding midpoint smoothing was applied to each coverage track, taking into account missing data by ignoring it and expanding the window size symmetrically around a midpoint to always include 100,000 datapoints, stepping the midpoint 10 kb each time.

Hominid Copy Number Analysis

Sequencing data for NA12878 (ERR194147), Vindjia Neanderthal (PRJEB21157), Altai Neanderthal (PRJEB1265), Denisovan (ERP001519), Chimpanzee (SRP012268), Gorilla (PRJEB2590) and Orangutan (SRR748005) were obtained either from SRA or from collaborators. These data were mapped to GRCh37 to obtain reads mapping to the *NOTCH2* (chr1:120,392,936-120,744,537) and *NOTCH2NL* (chr1:145,117,638-145,295,356) loci in that assembly, and then those reads were remapped to a reference containing just the GRCh38 version of *NOTCH2*. Coverage was extracted with bamCoverage, normalizing to 1x coverage across the custom *NOTCH2* reference. The resulting coverage tracks were then scaled to the average of the unique region of *NOTCH2* then underwent the same sliding midpoint normalization described above, with 5,000 datapoints per window and 2.5 kb step size.

Gordian Assembler

The extremely low number of long fragments per partition in the 10X Chromium process ensures that nearly all partitions containing sequence from a *NOTCH2NL* repeat will contain sequence from precisely one repeat copy. In order to recover the precise *NOTCH2NL* repeat sequences, a process was developed to assemble paratypes using barcoded reads. A 208 kb multiple sequence alignment of *NOTCH2NL* paralogs was constructed and a consensus sequence generated. For each sample being

assembled, the 10x Longranger pipeline was used to map enriched or unenriched reads to GRCh38. All reads that mapped to any of the five *NOTCH2* or *NOTCH2NL* loci in that alignment were extracted, as well as any reads associated with the same input molecules via the associated barcodes. These reads were then remapped to the consensus sequence. FreeBayes[73] [REF: https://arxiv.org/abs/1207.3907] was used to call variants on these alignments with ploidy set to 10 based on the putative number of NOTCH2NL repeats. Each barcode is then genotyped to find the set of alleles supported at each informative SNP site. Alleles for the majority of SNP sites are undetermined in each barcode due to the sparsity of the linked reads. The result is an MxB sparse matrix where M is the number of variants and B is the number of barcodes identified as having *NOTCH2*-like sequence. A statistical model is then used to phase this matrix into K paratypes. For each cluster of barcodes representing a single paratype, all reads with the associated barcodes are pooled for short-read assembly using the DeBruijn graph assembler idba_ud (Peng et al., 2012). This software is available on github at https://github.com/abishara/gordian_assembler.

Establishment of Paratypes in Population

The paratype assembly process described above was applied to the MYcroarray enriched 10x sequencing of NA19240, H9, NA12877, and the six Simons VIP samples. The H9 paratypes were validated with full-length cDNA sequencing (below). The NA12878/NA12891/NA12892 trio (Utah) as well as the NA24385/NA24143/NA24149 (Ashkenazi) trio were assembled using linked read data produced by 10x Genomics for the Genome In A Bottle Consortium. Inheritance was established for the Ashkenazi trio, as well as for the three NA12878 paratypes that assembled. Inherited paratypes are not double counted in **Table S1**. A scaffolding process using alignments of contigs to GRCh38 was performed to construct full-length *NOTCH2NL* loci for each of these assemblies. The *NOTCH2NL* transcripts were annotated and assessed for their protein level features.

Enrichment and Sequencing of Full-Length cDNA

Full-length cDNA was constructed from both week 5 cortical organoids as well as undifferentiated H9 hESC total RNA similar to previously described protocols (Byrne et al., 2017)] and were enriched using the same MyBaits oligonucleotide set as the 10x Chromium libraries. These cDNA libraries were prepared

and sequenced on the Oxford Nanopore MinION. 47,391 reads were obtained for the undifferentiated cells and 118,545 reads for the differentiated cells. The reads were base called with Metrichor. After pooling these datasets, the reads were aligned to GRCh38 to identify putative *NOTCH2NL* reads. 2,566 reads were identified in the week 5 dataset that mapped to *NOTCH2NL*, and 363 in the undifferentiated. Both datasets were filtered for full-length transcripts by requiring at least 70% coverage to the first 1.1 kb of the consensus sequence. This filtering process removed *NOTCH2* like transcripts, leaving a final set of 1,484 transcripts pooled across both timepoints to be analyzed.

Validating H9 Haplotypes Using Full-Length cDNA

The 1,484 *NOTCH2NL* transcript sequences identified above were aligned to a consensus sequence of H9 ESC transcript paratypes using MarginAlign (Jain et al., 2015). The reads were then reduced into feature vectors containing variant sites along the first 1.1 kb of the consensus to eliminate noise related to alternative transcription stop sites. The feature vectors were aligned using a Hidden Markov Model with one path for each of the paratype assemblies. Since the transcripts are already aligned to a consensus, there is no need for reverse transitions in the model, and since variation or recombination between paralogs is already accounted for in the assemblies, no transitions between paths are allowed. This vastly simplifies the Forward algorithm, and the maximum probability path (usually determined with the Viterbi Algorithm) is trivial to calculate under these conditions. All mismatches were assumed to be errors and were given an emission probability of 0.1 to approximate the error rate of the nanopore. The paratype assembly was validated by showing that there were no recurrent feature vectors that did not align well to any path through this model.

CRISPR Mutation of *NOTCH2NL* in the H9 ES Line

H9 hESC (WA09, WiCell Research Institute) at passage 42 were plated on a 6-well dish at 40-50% confluency. After 24 hours, cells were treated with 10μM ROCK inhibitor (Y27632; ATCC, ACS-3030) for 1 hour. 2.5μg of each guide plasmid (E2.1& E5.2, Fig. S5 cloned into pX458, Addgene) was then introduced for 4 hours using Xfect DNA transfection reagent (Clontech, 631317). Each guide set was introduced to all 6 wells of a 6-well plate. 48 hours after transfection, cells were dissociated from wells using Accutase cell dissociation enzyme (eBioscience, 00-4555-56), then rinsed twice in PBS

supplemented with 0.2mM EDTA, 2% KnockOut Serum Replacement (Thermo Fisher, 10828028), 1% Penicillin-Streptomycin (LifeTech, 15140122), and 2μM thiazovivin (Tocris, 1226056-71-8), and resuspended in a final volume of 1mL of sorting buffer. The cells were then filtered in a 70μm filter and sorted on a FACS Aria II (BD Biosciences) with a 100μm nozzle at 20psi to select for cells expressing the Cas9-2A-GFP encoded on pX458. Gating was optimized for specificity. Single cells positive for GFP were plated on a 10 cm plate containing 1.5x106 mouse embryonic fibroblasts (MEFs) and cultured in E8 Flex with 2μM thiazovivin (Tocris, 1226056-71-8) for added for the first 24 hours. After growing 5-7 days, individual colonies were manually isolated into 1 well of a 6-well dish on 250,000 MEFs in E8 Flex. 3-5 days later, 3-7 good colonies at passage 42+3 were frozen in BAMBANKER (Fisher Scientific, NC9582225). Remaining cells on MEFs were used for PCR deletion assay. For all subsequent analysis, cells were adapted to culturing on vitronectin (Thermo Fisher A14700) in Gibco's Essential 8 Flex medium (Thermo Fisher, A2858501).

PCR assay for CRIPSR deletion: For each clone, gDNA was isolated from one 70% confluent well of a 6-well dish using Zymo Quick-gDNA Miniprep kit (Zymo, D3006) according to the manufacturer's protocol. PCR was performed using approximately 70ng gDNA with Herculase II fusion DNA polymerase (Agilent, 6006745) using primers N2NL E2del_F (5' CACAGCCTTCCTCAAACAAA 3') and N2NL E5del_R (5' GTGCCACGCATAGTCTCTCA 3'). PCR products of the expected size were cloned and sequenced to determine that at least one of NOTCH2NL harbored the expected deletion. Positive clones were subject Chromium library preparation, target enrichment, Illumina sequencing and NOTCH2NL gene assembly as described above.

Estimate of *NOTCH2* and *NOTCH2NL* Expression in human fetal brain scRNA-Seq data

To asses *NOTCH2NL* expression in the developing brain, we re-analyzed single cell RNA sequencing data from (Nowakowski, et al, *Science*, In Press). Initial analysis of this data showed low expression of *NOTCH2* and *NOTCH2NL* presumably due to removal of multi-mapping reads. To address this, we constructed a custom Kallisto reference based off GENCODE V19 (hg19) where we removed the transcripts ENST00000468030.1, ENST00000344859.3 and ENST00000369340.3. The reads for 3,466

single cells were then quantified against this Kallisto index, and the *NOTCH2 and NOTCH2NL* rows of the resulting gene-cell matrix compared to previously generated tSNE clusters.

Copy Number Estimates of *NOTCH2NL* in Human Population

The copy number of *NOTCH2NLR* and *NOTCH2NLC* in the human population were established by extracting reads that map to *NOTCH2* (chr1:119,989,248-120,190,000), *NOTCH2NLR* (chr1:120,705,669-120,801,220), *NOTCH2NLA* (chr1:146,149,145-146,328,264), *NOTCH2NLB* (chr1:148,600,079-148,801,427) and *NOTCH2NLC* (chr1:149,328,818-149,471,561) from 266 individuals in the Simons Diversity Panel. These reads were then remapped to the 101,143 bp consensus sequence of a multiple sequence alignment of alignable portions of *NOTCH2* and all *NOTCH2NL* paralogs. This multiple sequence alignment was used to define our SUN markers, and the ratio of reads containing a SUN to a non-SUN were measured and the median value taken for *NOTCH2NLC* and *NOTCH2NLR*. Establishing copy number with SUNs proved difficult for *NOTCH2NLA* and *NOTCH2NLB* due to the high rate of segregating ectopic gene conversion alleles in the population. Each of the 266 samples was studied by hand. Using comparison to the 10 normal genomes assembled, it appeared that *NOTCH2NLA* and *NOTCH2NLB* are not copy number variable in the phenotypically normal population.

Paratype Estimation of *NOTCH2NL* in Human Population

Assigning paratypes without assemblies is not possible. To try and evaluate the gene conversion landscape in the population, we took the ratio of SUN read depths in all 266 Simons individuals as well as the six Simons VIP samples and our 10 assembled genomes and plotted them split up by paralog (**Table S4)**. These were evaluated for *NOTCH2NLC* and *NOTCH2NLR* copy number (**Figure S1G)**. Three samples were identified in Simons with apparent gene conversion in *NOTCH2NLC*, which we did not observe in any of our assembled samples. Manual analysis of these SUN diagrams led to the annotation of six distinct classes of *NOTCH2NLA-NOTCH2NLB* gene conversion with varying population frequencies. In some cases, the data were of lower quality and harder to interpret. The most common gene conversion allele is an overwrite of around 20kb of *NOTCH2NLB* by *NOTCH2NLA* in intronic sequence between exons 2 and 3, present in 42.5% of Simons normals haplotypes. When interpreting these SUN plots, it is helpful to remember that the denominator of the ratio is the total copy number, and

as such as individuals stray from N=10 the expected values change. Gene conversion can be observed as regions where one paralog has ratios on the Y axis go up while the other goes down. Exons 1-5 are located at 19,212-19,590 bp, 59,719-59,800 bp, 84,150-84,409 bp, 92,421-92,756 bp and 93,009-97,333 bp respectively in the consensus sequence.

Copy Number Estimates of Microarray 1q21.1 Deletion/Duplication Syndrome Patients

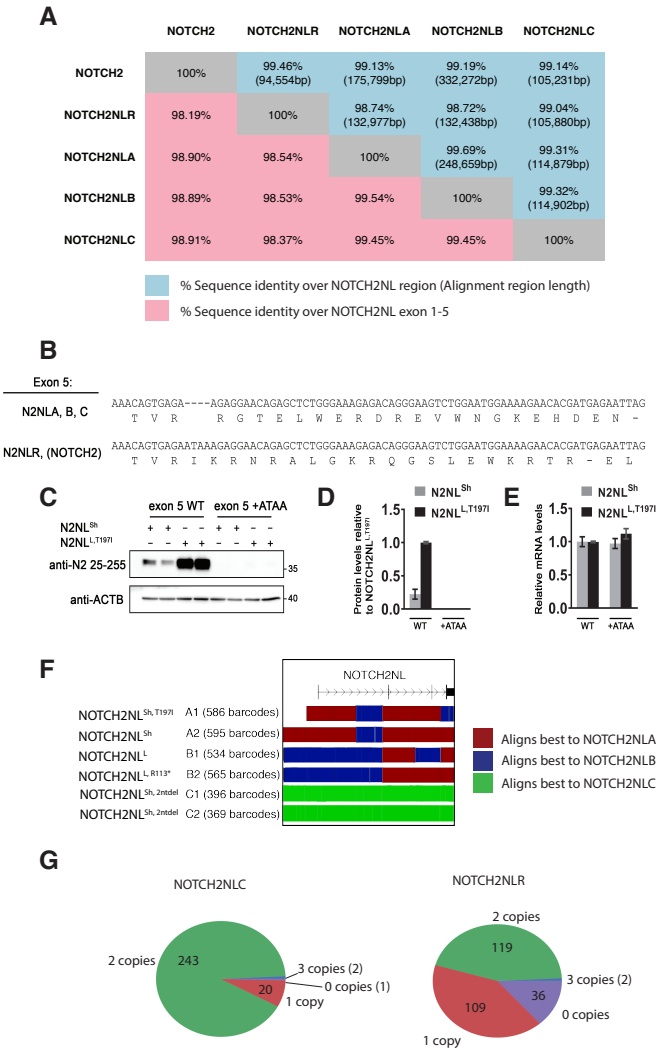
Comparative genomic hybridization (CGH) microarray probes from Agilent array designs #014693, #014950, #24616, and #16267 were mapped to the human hydatidiform mole genome assembly using the pblat aligner (https://code.google.com/p/pblat/). IDs for microcephaly and macrocephaly patients visualized in **Figure S7** are as follows: control-1: MCL08273; control-2: GSM1082800; 1: LG_252808110380_S01; 2: WSX002375; 3: SGM250214; 4: MCL08277; 5:MUG249341; 6: MCL04601; 7: MCL00270; 8: MCL02135; 9: NGS260131; 10: MCL01089; 11: MCL01415. The relative affinities of each probe for each of its mapping locations were calculated using the DECIPHER R package; the predicted hybridization efficiency of a probe and a mapping location was divided by the predicted hybridization efficiency of the probe and its reverse complement, yielding the affinity for the mapping location. An integer linear programming (ILP) model was created for each sample, with integral variables representing the copy number of each probe mapping location in the CGH sample genome and the CGH reference genome, and additional variables representing the total affinity-weighted copy number of all mapping locations for each probe. Approximate equality constraints were added to represent the measured sample/normal hybridization ratios (normalized per-probe to an average control value of 1.0 for arrays with control samples (GEO accession GSE44300) available), the chromosomal structure of the genome, and the prior belief that the copy number at most locations would be 2. Each approximate equality constraint was constructed by creating two variables, which were restricted to be positive and greater than the difference between the constrained quantities in each direction; the weighted sum of these two variables was then added into the model's objective function. In an alternate method, the locations of homologous sequences that could provide a plausible mechanism for duplication or deletion of a region were also added (data not shown). Finally, the resulting ILP model was minimized using

CBC[79] (https://projects.coin-or.org/) to produce the sample's integral copy number calls. The copy number inference software used is available from https://bitbucket.org/adam_novak/copynumber.

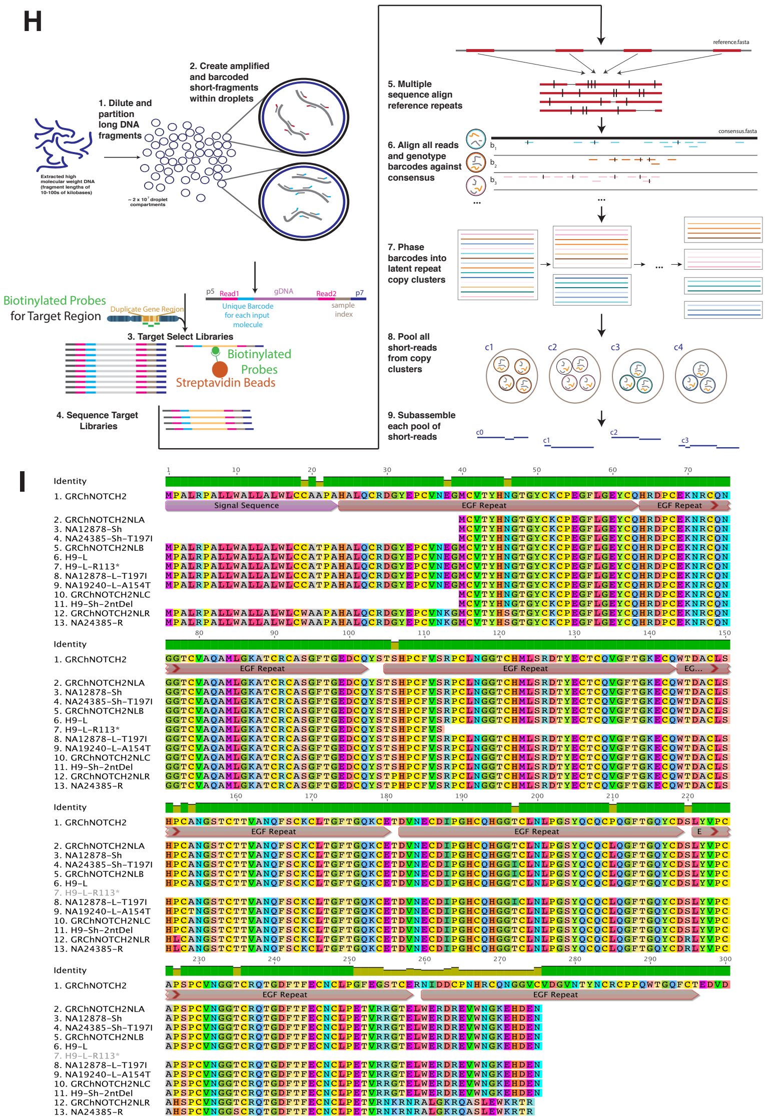
NOTCH2NL Expression in Week 5 Neurospheres

Two replicates of bulk RNA-seq of week 5 cortical organoids derived from H9 ES as well as undifferentiated cells from the H9 differentiation time course described above were quantified against a custom Kallisto reference based on GENCODE V27. Using bedtools, all transcripts which overlapped our curated annotations of *NOTCH2NL* paralogs and *NOTCH2* were removed. After converting this annotation set to FASTA, a subset of our paratype assemblies of H9 *NOTCH2NL* paralogs were added in. Only one representative of both *NOTCH2NLR* and *NOTCH2NLC* was used due to their high similarity on the transcript level. The TPM values of the replicates were averaged.

Figure S1


A
NOTCH2
NOTCH2NLR
NOTCH2NLA
NOTCH2NLB
NOTCH2NLC
100%
99.46% (94,554bp)
99.13% (175,799bp)
99.19% (332,272bp)
99.14% (105,231bp)
98.19%
100%
98.74% (132,977bp)
98.72% (132,438bp)
99.04% (105,880bp)
98.90%
98.54%
100%
99.69% (248,659bp)
99.31% (114,879bp)
98.89%
98.53%
99.54%
100%
99.32% (114,902bp)
98.91%
98.37%
99.45%
99.45%
100%
% Sequence identity over NOTCH2NL region (Alignment region length)
% Sequence identity over NOTCH2NL exon 1-5
B
Exon 5:
N2NLA, B, C
AAACAGTGAGA----AGAGGAACAGAGCTCTGGGAAAGAGACAGGGAAGTCTGGAATGGAAAAGAACACGATGAGAATTAG
T V R R G T E L W E R D R E V W N G K E H D E N -
N2NLR, (NOTCH2)
AAACAGTGAGAATAAAGAGGAACAGAGCTCTGGGAAAGAGACAGGGAAGTCTGGAATGGAAAAGAACACGATGAGAATTAG
T V R I K R N R A L G K R Q G S L E W K R T R - E L
C
exon 5 WT
exon 5 +ATAA
N2NLSh
N2NLL,T197I
anti-N2 25-255
anti-ACTB
35
40
D
N2NLSh
N2NLL,T197I
Protein levels relative to NOTCH2NLL,T197I
1.5
1.0
0.5
0.0
WT
+ATAA
E
N2NLSh
N2NLL,T197I
Relative mRNA levels
1.5
1.0
0.5
0.0
WT
+ATAA
F
NOTCH2NL
NOTCH2NLSh,T197I
NOTCH2NLSh
NOTCH2NLL
NOTCH2NLL,R113*
NOTCH2NLSh,2ntdel
NOTCH2NLSh,2ntdel
A1 (586 barcodes)
A2 (595 barcodes)
B1 (534 barcodes)
B2 (565 barcodes)
C1 (396 barcodes)
C2 (369 barcodes)
Aligns best to NOTCH2NLA
Aligns best to NOTCH2NLB
Aligns best to NOTCH2NLC
G
NOTCH2NLC
2 copies
243
3 copies (2)
0 copies (1)
20
1 copy
NOTCH2NLR
2 copies
119
3 copies (2)
36
0 copies
109
1 copy

H



1. Dilute and partition long DNA fragments
2. Create amplified and barcoded short-fragments within droplets
Biotinylated Probes for Target Region
Duplicate Gene Region
3. Target Select Libraries
Biotinylated Probes
Streptavidin Beads
4. Sequence Target Libraries
p5
Read1
gDNA
Read2
p7
Unique Barcode for each input molecule
sample index
5. Multiple sequence align reference repeats
reference.fasta
6. Align all reads and genotype barcodes against consensus
consensus.fasta
7. Phase barcodes into latent repeat copy clusters
8. Pool all short-reads from copy clusters
c1
c2
c3
c4
9. Subassemble each pool of short-reads
c0
Identity
1. GRChNOTCH2
Signal Sequence
EGF Repeat
2. GRChNOTCH2NLA
3. NA12878-Sh
4. NA24385-Sh-T197I
5. GRChNOTCH2NLB
6. H9-L
7. H9-L-R113*
8. NA12878-L-T197I
9. NA19240-L-A154T
10. GRChNOTCH2NLC
11. H9-Sh-2ntDel
12. GRChNOTCH2NLR
13. NA24385-R

I

Figure S1. Related to Figure 1. (A) Identity between NOTCH2-related genes as measured over the alignable genomic region (blue) or the gene exons (pink). (B) DNA and amino acid sequence of human NOTCH2NL genes in exon 5, which is derived from NOTCH2 intron 4. (C) Western blot of mouse ESCs transfected with WT NOTCH2NL, or NOTCH2NL with ancestral ATAA inserted in exon 5. (D) Relative protein levels based on western blot in panel C. (E) RT-qPCR analysis of the same samples for determination of transcript levels for each condition. (F) Alignment of the H9 assembled paratypes to GRCh38. Each paratype is colored as to whether a position aligns best to GRCH38 *NOTCH2NLA*, *NOTCH2NLB* or *NOTCH2NLC*. (G) Observed frequency of individuals with the indicated NOTCH2NLC and NOTCH2NLR copy number in the Simons Diversity Panel (n=266). (H) Schematic of linked-read sequencing and Gordian Assembler protocol using the 10X genomics Chromium genome assay and oligo-capture to enrich for library fragments containing the desired genomic region. (I) Protein alignment of observed NOTCH2, NOTCH2NL and NOTCH2NLR paratypes based on our assembly results. Note that NOTCH2 sequence extends beyond what is shown in the alignment. A segregating variant in NOTCH2NLR is found at amino acid position 235.

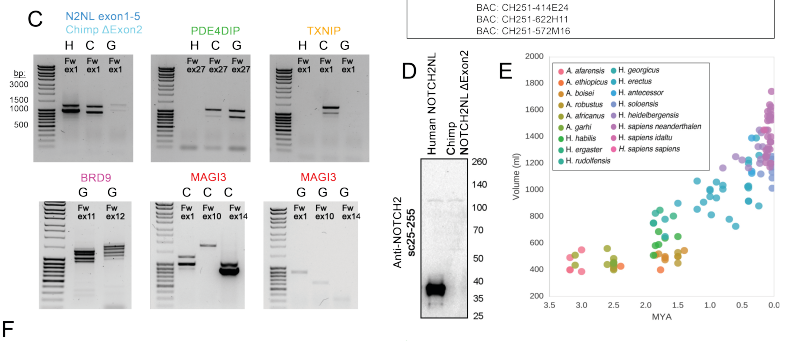
Figure S2

A

Gorilla	Genome support	Breakpoint coordinates hg38 (BLAT)
PDE4DIP[1-27]	BAC: CH277-26C18 TI: 1679409700 TI: 2019116386 TI: 1680961471 TI: 2192386756	chr1:149,005,377 - chr1:149,426,133
MAGI3[1-14]	TI: 1671716786 TI: 1671781225 TI: 2033420236	chr1:113,651,629 - chr1:146,175,285
BRD9[1-12]	BAC: CH277-57H6	chr5:872,955 - chr1:147,072,407, chr1:147,069,489 - chr1:148,619,452

B

Chimp	Genome support	Breakpoint coordinates hg38 (BLAT)
PDE4DIP[1-27]	WGS: AACZ03011197 WGS: AADA01102044 WGS: AACZ03149689	chr1:149,005,377 - chr1:149,426,133
MAGI3[1-14]	WGS: AACZ03011046 WGS: AADA01172734	chr1:113,651,659 - chr1:146,175,285
MAGI3[1]	WGS: AACZ03149796 WGS: AADA01170825	chr1:113,536,139 - chr1:146,169,274
TXNIP[1]	WGS: AACZ03149820	chr1:145,996,530 - chr1:146,226,704
ΔExon2	BAC: CH251-305M22 BAC: CH251-243D3 BAC: CH251-414E24 BAC: CH251-622H11 BAC: CH251-572M16	chr1:148,675,676 - chr1:148,623,447



F

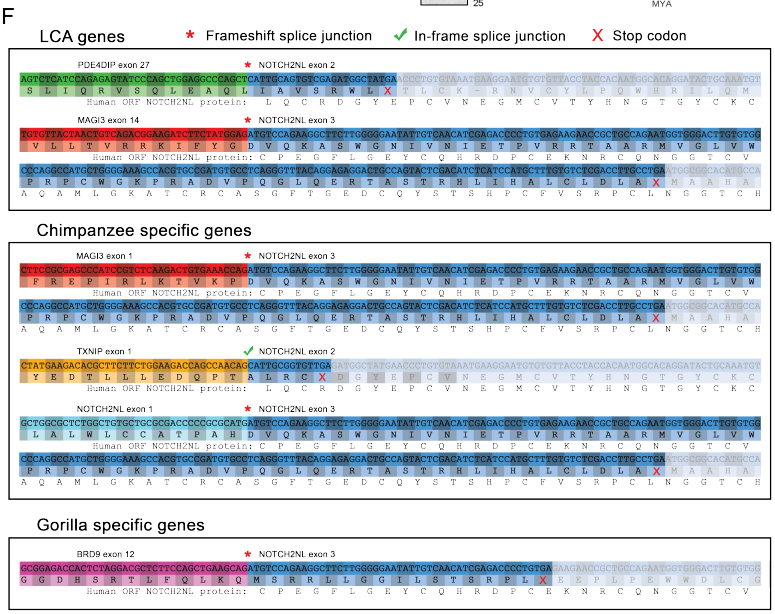


Figure S2. Related to Figure 2. Details of NOTCH2NL-like genes in Gorilla (A) and Chimp (B) and source of genome sequence support. (C) Verification of fusion genes by RT-PCR on Chimp and Gorilla RNA. H = Human, C = Chimpanzee, G = Gorilla. (D) Immunoblot blot using an N-terminal NOTCH2 antibody (aa 25-255), comparing ectopic expression of Chimp NOTCH2NL-like gene lacking exon 2 compared to ectopic expression of human *NOTCH2NL*^{Sh} in mESCs. (E) Endocranial volume of fossil hominids versus time as determined by Holloway, et al. 2004. (F) Details of splice junctions of fusion genes and related open reading frames. Top rows show nucleotide sequence in fusion transcripts. Middle rows show peptides derived from these transcripts. Lower rows contain the orthologous human NOTCH2NL protein sequence.

Figure S3

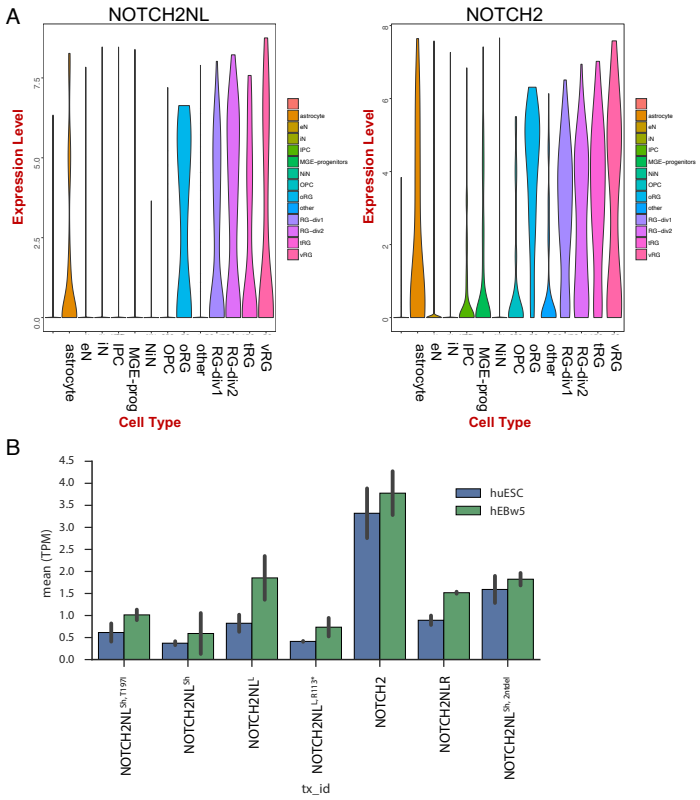


Figure S3. Related to Figure 3. (A) Summary violin plots indicating *NOTCH2NL* and *NOTCH2* expression in various cell types. (B) *NOTCH2NL* paratype expression in undifferentiated hESCs and week 5 cortical organoids from bulk Illumina RNA-Seq.

Figure S4

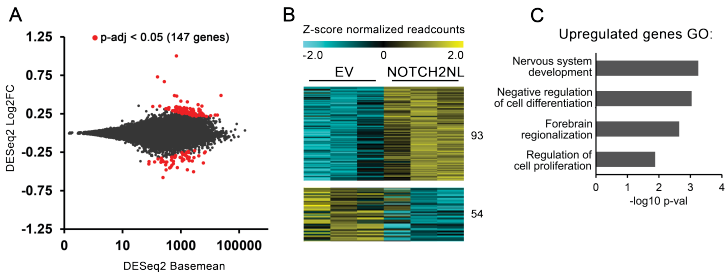


Figure S4. Related to Figure 4. (A) MA plot of RNA-sequencing data of mouse cortical organoids based on DESeq2 analysis. (B) Z-score of differentially expressed genes (p-adj < 0.05, DESeq2). (C) GO terms significantly associated with the upregulated genes in organoids ectopically expressing NOTCH2NL$^{Sh.T197I}$.

Figure S5

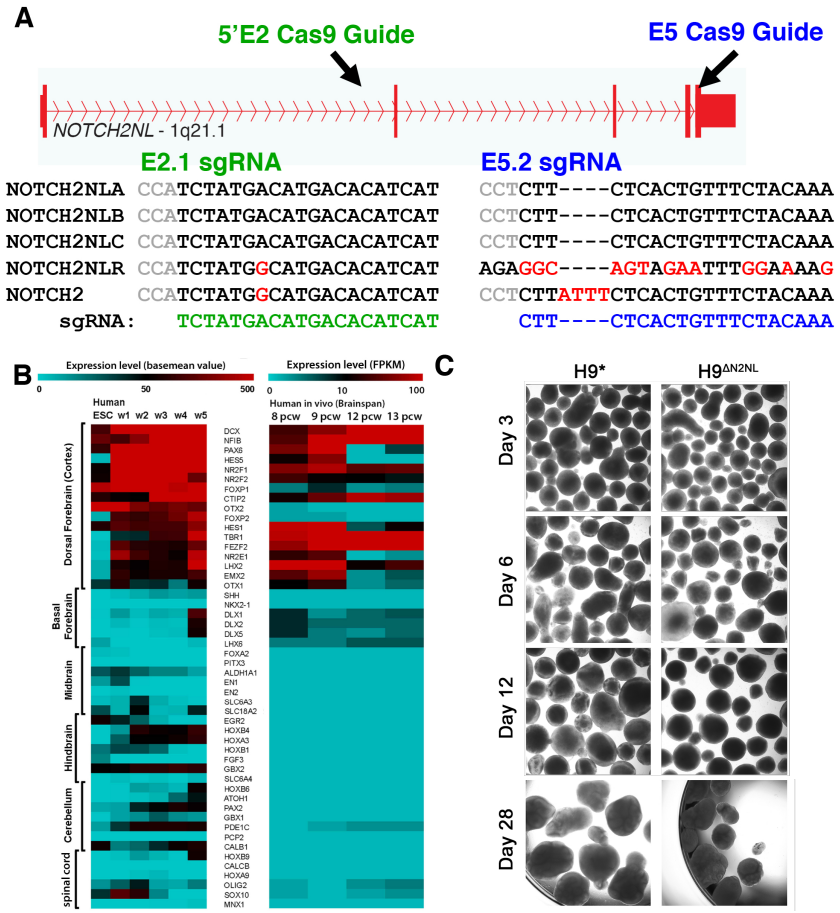


Figure S5. Related to Figure 5. (A) Schematic of strategy to generate *NOTCH2NL*-specific deletions using CRISPR/Cas9 and alignment of the two guide sequences used to NOTCH2NL-related sequences. (B) Heatmap of expression levels for a selection of brain structure marker genes from hESC-derived cortical organoids at the indicated time points (Left) and from human embryonic dorsal prefrontal cortex (DFC) samples at 8 pcw, 9 pcw, 12 pcw and 13 pcw, derived from the Allen Brain Atlas (http://www.brainspan.org). w = week; pcw = post conception week; FPKM= fragments per kilobase of exon per million fragments mapped. (C) Brightfield images of developing H9* and H9$^{NOTCH2NL\Delta}$ cortical organoids.

Figure S6

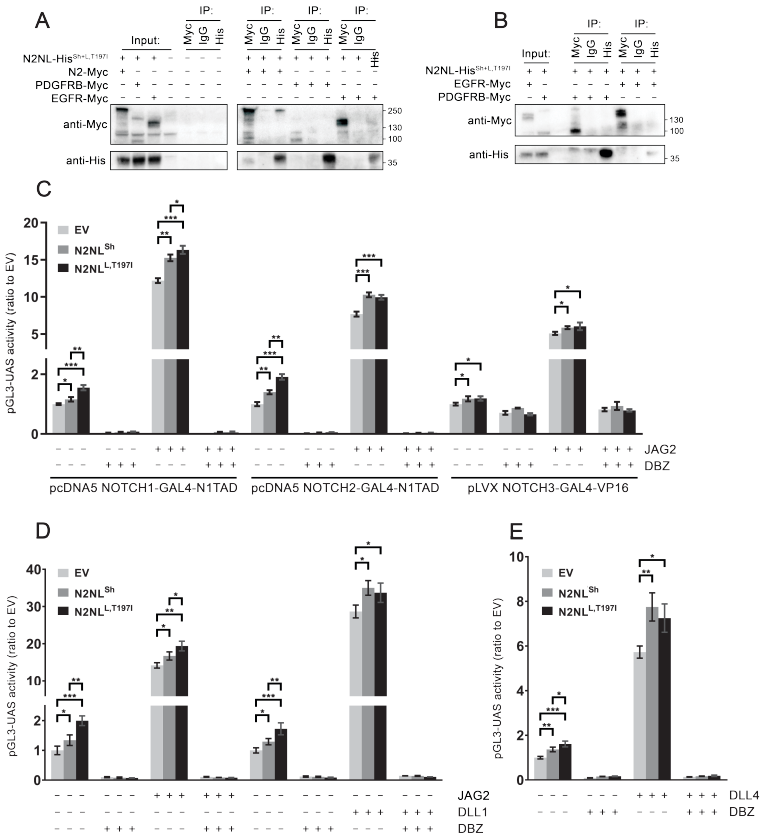


Figure S6. Related to Figure 6. (A-B) Investigation of co-immunoprecipitation of NOTCH2NL with PDGFRB and EGFR in two independent experiments. (C) Reporter assay to assess the effect of NOTCH2NL using either NOTCH1-GAL4, NOTCH2-GAL4 or NOTCH3-GAL4 to induce pGL3-UAS reporter activation. 6 replicates in one experiment. Student's t-test with Holm-Bonferroni correction (* $p < 0.05$, ** $p < 10^{-3}$, *** $p < 10^{-5}$), error bars indicate SD. (D) NOTCH2NL effects of NOTCH2-GAL4 reporter assay co-culture with JAG2 or DLL1 expressing cells. 6 replicates in one experiment. Student's t-test with Holm-Bonferroni correction (* $p<0.05$, ** $p<10^{-3}$, *** $p<10^{-5}$), error bars indicate SD. (E) NOTCH2NL remains effective in an assay with activation of NOTCH2-GAL4 using recombinant DLL4 coated plates. Average of 3 independent experiments with 4 or 5 replicates each. Two-way anova with Tukey's HSD (* $p < 10^{-4}$, ** $p <10^{-8}$, *** $p < 10^{-12}$), error bars indicate SEM.

Figure S7

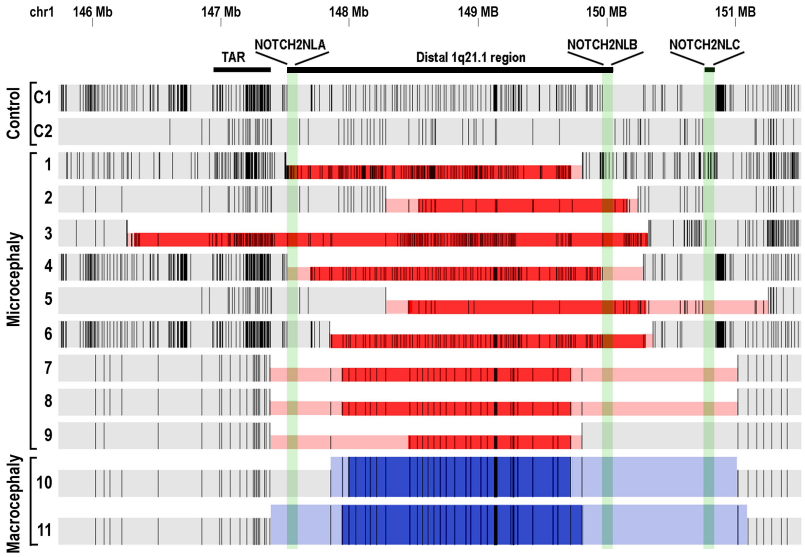


Figure S7. Related to Figure 7. Relative probe intensities from CNV-microarrays for 2 controls and 11 patients with reported 1q21.1 aberrations mapped to the GRCh38 1q21.1 assembly. Gray: normal, red: deletion, blue: duplication. Dark red/blue is high confidence deletion/duplication based on probe values, light red/blue are potentially part of the deletion/duplication.

Table S1. Related to Figures 1. Results of NOTCH2NL gene de novo assembly.

		NOTCH2NLA/NOTCH2NLB paratypes								NOTCH2NLC	
Sample	Type	Sh	Sh-T197I	Sh-A154T	L	L-T197I	L-R113*	L-A154T, R113*	Ambiguous A/B paratype	Sh-2ntDel	Total A/B/C copies
CHM1	Normal	2				2				2	6
H9	Normal	1	1		1			1		2	6
NA12878	Normal	1		1		1			1	2	6
NA12891	Normal	2	1			1				1	5
NA12892	Normal	1	1		1				1	2	6
NA19240	Normal	2	1		1					2	6
NA24143	Normal	2	2							1	5
NA24149	Normal	3				1				2	6
NA24385	Normal	2	1			1				1	5
SV721	Simons VIP	1	2		1		1			2	7
SV735	Simons VIP	2	1							2	5
SV770	Simons VIP	1			1	1				2	5
SV780	Simons VIP	1	1		1					2	5
SV788	Simons VIP	1	1						1	2	5
SV877	Simons VIP	2	1							2	5

Table S2. Related to Figure 4 and 5. Gene expression measurements from RNA-seq experiments.

Sup_Table_2_RNA-Seq_GeneExpression_V1.xlsx supplied as an Excel Workbook

Table S3. Related to Figure 7. Features of Simons VIP samples.

Sample	SFARI ID	Sex	1q21.1 Genetic Status	Inheritance	NOTCH2NL A/B Genotype	NOTCH2NL C Genotype	Autism Spectrum Disorder	Intellectual Disability/ Developmental Delay	Other
SV721	14815-x13	M	duplication	inherited	L, Sh, L-R113*, Sh-T197I, Sh-T197I	Sh-2ntdel x2	Y	N	ADHD
SV735	14836-x45	F	deletion	inherited	Sh, Sh, Sh-T197I	Sh-2ntdel x2	N	Y	
SV770	14843-x15	F	deletion	de-novo	L, L-T197I, Sh	Sh-2ntdel x2	N	Y	Speech Sound Disorder
SV788	14842-x1	F	deletion	unknown	Sh, Sh-T197I, ?	Sh-2ntdel x2	N	N	Anexiety, Developmental Coordination Disorder
SV780	14816-x12	M	deletion	inherited	L, Sh, Sh-T197I	Sh-2ntdel x2	N	N	Tourette's
SV877	14879-x8	M	deletion	inherited	Sh, Sh, L-T197I	Sh-2ntdel x2	Y	N	ADHD, Learning Disability

Table S4. Related to Figures 1 and 7. Curated Paratypes of Assembled and Simons Normals

Sup_Table_4_ParatypeCalls.xlsx supplied as an Excel Workbook

Chapter 7

Discussion

7.1 Introduction

Taken as a whole, this thesis presents a deep analysis of the field of comparative genomics. I have developed a novel comparative genomics framework that will be hugely useful to the field going forward into new eras of genome assembly and alignment.

7.2 Chapters

In chapter 1, I introduced the history of comparative genomics. As a field, comparative genomics saw its biggest explosion of analysis and publications in the early 2000s. As the era of short read sequencing began, the field languished due to a lack of high quality genomes providing avenues for new analysis. Now, in the era of long read assembly, comparative genomics is poised to become an important field for analyzing not just evolutionary history but also human diseases.

In chapter 2 I introduced the history of the mouse genomes project, a pivotal project showing the culmination of the short read assembly era, augmented by comparative techniques to sidestep shortcomings of short read assembly. Despite being begun over three years ago, only now are other projects like the 200 Mammals Project beginning to reach this scale of genome assembly.

In chapter 3 I introduced CAT, the comparative genomics framework I developed to leverage existing high quality annotation sets and whole genome alignments to

produce annotation sets and genome browsers that enable novel comparative genomics research to be performed.

In chapter 4 I included the entirety of the CAT publication that is in submission. This paper lays out the incredible usefulness of CAT, showing that it can improve annotations even on well studied genomes like Rat, and can outperform large institutional annotation pipelines like those at Ensembl and NCBI by leveraging high quality annotation sets like the carefully curated set that GENCODE produces for mouse and human.

In chapter 5 I provide a brief overview of the myraid projects that CAT has been used for over the years. The number of publications CAT has been involved with is a testament to the usefulness of the pipeline. Much thought has been given over the past few years to genome assembly, but less to alignmenet and annotation. As the number of high quality assembled genomes grows, and the assemblies are increasingly produced by individual labs instead of large consortia, a tool like CAT is required to automated annotation in a fashion usable by individual labs and bioinformatics scientists. Additionally, the ability of CAT to track orthology relationships is unique in the realm of genome annotation.

In chapter 6 I changed gears and introduced the NOTCH2NL project. This project is also a comparative genomics project at its heart, but also involves a lot of advanced molecular biology techniques. I did not personally perform any molecular biology experiments for the project, but I did design or help design most of the genomics-related experiments performed. The NOTCH2NL project was an opportunity to apply software development skills I learned while working on the CAT project to an extremely difficult region of the human genome.

7.3 Conclusion

This thesis presents multiple novel approaches in the comparative genomics space. This field is poised to make major advances in the upcoming years due to improvements in sequencing technology and falling costs. Projects like Genome 10K have been proposed for years, but their lofty goals are only becoming tractable now. Additionally, projects like the NOTCH2NL project are also becoming tractable through

advances in long read sequencing, particularly of transcriptomes. Methods like the enrichment process developed for both DNA and RNA in the NOTCH2NL project allow for specific regions inaccessible to standard methods to be analyzed at reasonable cost. These techniques provide avenues for research into human disease phenotypes. In the same sense, CAT provides this opportunity on a whole genome scale when run on individual human genome assemblies.

In the future, I hope CAT will be incorporated into large institutional pipelines like those at Ensembl. It seems clear that this novel approach to genome annotation provides new opportunity for annotation completeness as well as defining orthology relationships. My work shows that many of the alternative isoforms discovered through intensive curation like GENCODE are conserved across evolution and can be annotated comparatively. CAT will also play a role in the Vertebrate Genomes Project, the 200 Mammals Project, and Genome 10K as a whole. As mentioned in the CAT paper, there is room for these concepts to be taken further by applying machine learning techniques to the currently relatively simple consensus finding process. Additionally, more work can be done to completely annotate many-to-many orthology relationships.

Chapter 8

Supplementary Text

8.1 CAT README

Below I have included the CAT README file as of commit 79bc4b3. This provides a detailed description of the CAT pipeline, how it functions, and how it can be parameterized. I was unable to get the figures to import properly, but they are visible as figures 2.1 and 2.3 respectively.

This project aims to provide a straightforward end-to-end pipeline that takes as input a HAL-format multiple whole genome alignment as well as a GFF3 file representing annotations on one high quality assembly in the HAL alignment, and produces a output GFF3 annotation on all target genomes chosen.

This pipeline is capable of running both on local cluster hardware as well as on common cloud infrastructure using the toil (http://toil.readthedocs.io/en/latest/) workflow engine. For full runs on many genomes, a decent amount of computational effort is required. Memory usage is moderate.

Above is a flowchart schematic of the functionality of the `CAT` pipeline.

Installation

The pipeline can be installed by a simple `pip` install:

```
pip install git+https://github.com/ComparativeGenomicsToolkit/Comparative-Annotation-Toolkit.git
```

However, at this time, direct pip installation will mean that the `luigi.cfg`, `logging.cfg`, and test files will be buried in your python directory. I am still trying to figure out how to approach this problem. In the meantime, you may be better off instead cloning the directory and installing from your clone:

```
git clone git+https://github.com/ComparativeGenomicsToolkit/Comparative-Annotation-Toolkit.git
pip install -e Comparative-Annotation-Toolkit
```

If you want to do the direct pip installation, you can grab the config files from the repository and place them in whatever directory you want to execute from, or set the `LUIGI_CONFIG_PATH` environmental variable to point to their location. Or have an ugly log, your choice.

Either form of `pip` installation will install all of the python dependencies. However, there are binary dependencies that must be compiled and installed in addition.

Dependencies

1. Kent toolkit (https://github.com/ucscGenomeBrowser/kent) . Follow the installation instructions there. Make sure you put the newly created `~/bin/$MACHTYPE` directory on your path.
2. bedtools (http://bedtools.readthedocs.io/en/latest/) .

3. samtools (http://www.htslib.org/) (1.3 or greater).

4. Augustus (http://bioinf.uni-greifswald.de/augustus/binaries/) . Make sure you are installing `augustus >= 3.3`. You need to follow the instructions to compile `augustus` in comparative augustus mode. This requires that you modify a few lines in the `common.mk` file, and also need to have `sqlite3`, `lp-solve`, `bamtools`, and `libboost` installed. If you are using ubuntu, this should work: `apt-get install libboost-all-dev libboost sqlite3 libsqlite3-0 libsqlite3-dev libgsl0-dev lp-solve liblpsolve55-dev bamtools libbamtools-dev`

   After you have the primary `augustus` binaries compiled, add the directory to your path. Note that if you move the `augustus` binaries from their original location, you will need to set the `AUGUSTUS_CONFIG_PATH` global variable to point to the species directory.

   You will also need to put the contents of the `scripts` directory on your path. Next, you need to compile the following auxiliary programs from the folder `auxprogs`:

   1. `joingenes`. Compiling this program will place it in the `augustus` binary directory.
   2. `bam2hints`. Compiling this program will place it in the `augustus` binary directory. Requires `bamtools` to be installed. If the `bamtools` headers are not at `/usr/include/bamtools`, you will need to modify the makefile.
   3. `filterBam`. Also requires the `bamtools` headers.
   4. `bam2wig`. Compiling this program will NOT place it in the `augustus` binary directory, you must do so yourself. This program requires you modify the makefile to explicitly point to your installation of `htslib`, `bcftools`, `samtools`, and `tabix`. `Tabix` is now packaged with `htslib`, and both are included in your `kent` directory at `$kent/src/htslib/`.
   5. `homGeneMapping`. This program must also have its makefile at `$augustus/trunks/auxprogs/homGeneMapping/src/Makefile` modified to turn on the `BOOST = true` and `SQLITE = true` flags. Then run `make clean && make` to recompile.
   6. There are a series of perl scripts that you need to place on your path from the `$augustus/trunks/scripts` directory: `wig2hints.pl`, `blat2hints.pl`, `transMap2hints.pl`, and `join_mult_hints.pl`.

5. HAL toolkit (https://github.com/glennhickey/hal) . To install the HAL toolkit, you must also have the sonLib (https://github.com/benedictpaten/sonLib) repository in the same parent directory. Compile sonLib first, then compile hal. Once hal is compiled, you need to have the binaries on your path.
6. wiggletools (https://github.com/Ensembl/WiggleTools) . Used to combine RNA-seq expression in assembly hubs.
7. sambamba (https://github.com/lomereiter/sambamba/releases) . Used to name sort faster than samtools for hints building.

In total, you must have all of the binaries and scripts listed below on your path. The pipeline will check for them before executing steps. `hal2fasta halStats halLiftover faToTwoBit pyfasta gff3ToGenePred genePredToBed genePredToFakePsl bamToPsl blat2hints.pl gff3ToGenePred join_mult_hints.pl pslPosTarget axtChain chainMergeSort pslMap pslRecalcMatch pslMapPostChain augustus transMap2hints.pl joingenes hal2maf`

```
gtfToGenePred genePredToGtf bedtools homGeneMapping blat pslCheck pslCDnaFilter
pslToBigPsl bedSort bedToBigBed sambamba wig2hints.pl
```

Running the pipeline

This pipeline makes use of Luigi (https://github.com/spotify/luigi) to link the various steps together. First, start the `luigid` daemon:

```
luigid --background --logdir luigi_logs
```

Which provides the central scheduler as well as the web UI, which can be accessed at `localhost:8082`. If you don't want to use the daemon, add the flag `--local-scheduler` to the invocation.

To run the test data, change directories to the CAT installation folder and do the following:

```
luigi --module cat RunCat --hal=test_data/vertebrates.hal --ref-genome=mm10 --
workers=10 --config=test_data/test.config --work-dir test_install --out-dir
test_install --local-scheduler --augustus --augustus-cgp --augustus-pb --assembly-hub
> log.txt
```

The test should take around 30 minutes to execute. You can track progress in the log file.

Command line options

As described above, the primary method to executing the pipeline is to follow the invocation `luigi --module cat RunCat --hal=${halfile} --ref-genome=${ref-genome} --config=${config}`. Below are the flags that can modify execution and output.

Main options

`--hal`: Input HAL alignment file. (REQUIRED).

`--ref-genome`: Reference genome sequence name. Must be present in HAL. (REQUIRED).

`--config`: Path to the config file with annotations and extrinsic hints. See the config section for more information. (REQUIRED).

`--out-dir`: Output directory. Defaults to `./cat_output`.

`--work-dir`: Working directory. Defaults to `./cat_work`. Stores all the intermediate files as well as the `toil` jobStore. Can be removed after completion (but not if you want to re-do any steps).

`--target-genomes`: List of genomes to use. If not set, all non-reference genomes in the HAL are used. Due to how `luigi` handles command line tuple parameters, this flag must be formatted as if it was a tuple being passed directly to python, single quoted. So, for example, if

your target genomes were Human and Mouse, then you would pass `--target-genomes='("Human", "Mouse")'`. As always with python tuples, if you have only one member, you must have a trailing comma.

`--workers`: Number of local cores to use. If running `toil` in singleMachine mode, care must be taken with the balance of this value and the `--maxCores` parameter.

transMap options

`--local-near-best`: Adjusts the `localNearBest` parameter passed to `pslCDnaFilter`. This algorithm attempts to resolve paralogous alignments for a given source transcript, allowing for distinct portions of the transcript to align to different locations. The default value is 0.15, which is a fairly lenient setting. Decreasing this value increases the number of alignments called as paralogous. If your assembly is fragmented, keep this value higher. Note that this parameter does not affect paralog resolution, just the calling of paralogous alignments (populates the Paralogy tag in the output).

`--minimum-paralog-coverage`: After the `localNearBest` algorithm is used to filter out paralogous alignments, the filtered alignments are evaluated for having at least this much coverage of the whole transcript before being considered a paralog. Default value is 25. Increasing this value restricts paralog calls.

AugustusTM(R) options

`--augustus`: Run AugustusTM(R)?

`--augustus-species`: What Augustus species do we want to use? If your species is not a mammal, please choose one of the species listed here (http://bioinf.uni-greifswald.de/augustus/) .

`--augustus-utr-off`: AugustusTMR will crash trying to predict UTRs if your `--augustus-species` lacks a trained UTR model. You can check if `$augustusDir/config/species/$augustusSecies/$augustusSpecies_utr_probs.pbl` exists. If it does not, set this flag.

AugustusCGP options

`--augustus-cgp`: Run AugustusCGP?

`--cgp-param`: Parameters file after training CGP on the alignment. See the AugustusCGP section.

`--maf-chunksize`: Size to chunk HAL into. Larger values make the CGP jobs take longer, but reduce problems related to splitting in genic regions. Default is 2500000. If your HAL contains more than 10 or so genomes, reducing this value to 1000000 or so is a good idea to keep job run-times below an hour and avoid going over 8GB of RAM per job. For a 25-way alignment, I set this value to 750000.

`--maf-overlap`: How much overlap to use in HAL chunks. Larger values increase redundant predictions (which are merged). Default is 500000. For a 25-way alignment, I set this value to

150000.

`--cgp-train-num-exons`: Number of exons to require in the alignment subset used for training CGP. See the AugustusCGP section. Default is 5000.

AugustusPB options

`--augustus-pb`: Run AugustusPB? Will only run on genomes with IsoSeq data in the config file.

`--pb-genome-chunksize`: Size to chunk genome into. Default is 20000000.

`--maf-overlap`: How much overlap to use in genome chunks. Default is 500000.

Filtering and consensus finding options

`--intron-rnaseq-support`: Amount of RNA-seq intron support a transcript must have to be considered. Must be a value between 0 and 100. Default is 0.

`--exon-rnaseq-support`: Amount of RNA-seq exon support a transcript must have to be considered. Must be a value between 0 and 100. Default is 0.

`--intron-annot-support`: Amount of reference intron annotation support a transcript must have to be considered. Must be a value between 0 and 100. Default is 0.

`--exon-annot-support`: Amount of reference exon annotation support a transcript must have to be considered. Must be a value between 0 and 100. Default is 0.

`--original-intron-support`: Amount of original intron support. See transcript evaluation description of original introns a transcript must have to be considered. Must be a value between 0 and 100. Default is 0.

`--denovo-num-introns`: For de-novo predictions, discard any transcripts with fewer than these number of introns. Important when RNA-seq data are noisy. Default is 0.

`--denovo-splice-support`: For de-novo predictions, discard any transcripts with less than this percent of RNA-seq intron support. Must be a value between 0 and 100. Default is 0.

`--denovo-exon-support`: For de-novo predictions, discard any transcripts with less than this percent of RNA-seq exon support. Must be a value between 0 and 100. Default is 0.

`--require-pacbio-support`: If set, all isoforms in the final set must be supported by at least one IsoSeq read. This flag is likely to discard a ton of transcripts, so be careful.

`--in-species-rna-support-only`: If set, all of the above intron/exon support flags will look only at RNA-seq/IsoSeq data from the species in question, and not make use of `homGeneMapping` to check support in all species. The output plots will always report in-species support.

`--rebuild-consensus`: A convenience flag to allow you to adjust the flags above. When set, will force the pipeline to re-run consensus finding and will also re-build the downstream plots and assembly hub.

See below for `toil` options shared with the hints database pipeline.

Toil

The remaining options are passed directly along to `toil`:

`--batchSystem`: Batch system to use. Defaults to singleMachine. If running in singleMachine mode, no cluster jobs will be submitted. In addition, care must be taken to balance the `--maxCores` field with the `--workers` field with the toil resources in `luigi.cfg`. Basically, you want to make sure that your # of toil resources multiplied by your `--maxCores` is fewer than the total number of system cores you want to use. However, I **highly** recommend using a non-local batch system. See the toil documentation for more.

`--maxCores`: The number of cores each `toil` module will use. If submitting to a batch system, this limits the number of concurrent submissions.

Config file

The config file contains the paths to two important pieces of information -- the reference GFF3 and the extrinsic hints (bams).

A major component of producing high quality comparative annotations is making use of RNA-seq and/or IsoSeq information. This information is used as hints to the `augustus` gene finding tool along with `transMap`, and is a major component of cleaning up transcript projections. This is also useful if you run the `augustusCGP` or `augustusPB` portions of the pipeline.

If the genetic distances in your alignment are high (say maybe an average identity in the 70s-80s), then you may derive great benefit from using a protein reference, if possible. This will be particularly useful for `augustusCGP`.

A template for the config file is below. At a minimum, your config file must have the annotation section. A example config file is provided in the `test_data` folder.

BAM files must be indexed!

```
[ANNOTATION]
Genome = /path/to/reference/gff3

[BAM]
Genome = /path/to/fofn <OR> /path/to/bam1.bam, /path/to/bam2.bam

[INTRONBAM]
Genome = /path/to/fofn/of/noisy/rnaseq

[ISO_SEQ_BAM]
Genome = /path/to/isoseq/bams
```

```
[PROTEIN_FASTA]
Genome = /path/to/protein/fasta
```

Note that the BAM/INTRONBAM/ISO_SEQ_BAM fields can be populated either with a comma separated list of BAMs or a single file with a line pointing to each BAM (a FOFN, or file-of-file-names). The reference sequence information will be extracted from the HAL alignment.

For the PROTEIN_FASTA field, every genome you wish to have the protein fasta be aligned to must be on its own separate line. All of these can point to the same FASTA.

RNA-seq libraries

It is **extremely** important that you use high quality RNA-seq. Libraries should be poly-A selected and paired end with a minimum read length of 75bp. If any of these are not true, it is advisable to place these libraries in the INTRONBAM field. Any genome can have a mix of BAM and INTRONBAM hints.

BAM files must be indexed!

ISoSeq libraries

If you are using IsoSeq data, it is recommended that you doing your mapping with `gmap`. Follow the tutorial (https://github.com/PacificBiosciences/cDNA_primer/wiki/Aligner-tutorial:-GMAP,-STAR,-BLAT,-and-BLASR) .

GFF3 Reference

CAT relies on a proper GFF3 file from the reference. One very important part of this GFF3 file is the `biotype` tag, which follows the GENCODE/Ensembl convention. The concept of a `protein_coding` biotype is hard baked into the pipeline. Proper division of biotypes is very important for transMap filtering and consensus finding to work properly.

If your GFF3 has duplicate transcript names, the pipeline will complain. One common cause of this is PAR locus genes. You will want to remove PAR genes -- If your GFF3 came from GENCODE, you should be able to do this: `grep -v PAR $gff > $gff.fixed`

Execution modes

The default mode of this pipeline will perform the following tasks:

1. Lift all annotations present in the input GFF3 to all other genomes in the alignment.
2. Filter these comparative transcripts for paralogous mappings.
3. Evaluate these transcripts for potential problems, assigning a score.

4. Produce a output annotation set as well a series of plots charting how this process went.

These steps will run reasonably fast on one machine without any need for cluster computing. However, to construct a high quality annotation set, it is recommended that the pipeline be run with as many modes of AUGUSTUS as possible.

AugustusTM(R)

The primary parameterization of AUGUSTUS for comparative annotation is primarily a method to clean up transMap projections. Due to a combination of assembly error, alignment noise and real biological changes transMap projections have frame shifting indels, missing or incomplete exons, and invalid splice sites. AugustusTM is given every protein coding transMap projection one at a time with some flanking sequence and asked to construct a transcript that closely matches the intron-exon structure that transMap provides. Since AUGUSTUS enforces a standard gene model, frame shifts and invalid splices will be adjusted to a valid form. In some cases this will mangle the transcript, producing either another isoform or something that does not resemble the source transcript. AugustusTMR runs the same genomic interval and transMap derived hints through AUGUSTUS a second time, but with less strict weights on the transMap hints and with the addition of extrinsic hints from RNA-seq and/or IsoSeq. This is particularly useful in regions where an exon was dropped in the Cactus alignment.

AugustusTM and AugustusTMR can be ran by providing the `--augustus` flag to the pipeline. AugustusTMR will only be ran for genomes with extrinsic information in the hints database. If you are running CAT on a non-mammal, you will want to modify the `--augustus-species` flag to one of the species listed here (http://bioinf.uni-greifswald.de/augustus/) . Take care to check if your species has a UTR model, and adjust the `--augustus-utr-off` flag accordingly.

AugustusCGP

augustusCGP is the comparative mode of AUGUSTUS recently introduced by Stefanie Nachtweide (https://academic.oup.com/bioinformatics/article/32/22/3388/2525611/Simultaneous-gene-finding-in-multiple-genomes) . This mode of AUGUSTUS takes as input a HAL format multiple whole genome alignment and simultaneously produces ***ab-initio*** transcript predictions in all genomes, taking into account conservation as well as any extrinsic information provided. AugustusCGP allows for the introduction of novel isoforms and loci in the final gene sets.

AugustusCGP can be ran by providing the `--augustus-cgp` flag to the pipeline. If no previously trained model is provided to AugustusCGP via the `--cgp-param` flag, then the pipeline will automatically train the model using the given alignment. To do so, random subsets of the alignment will be extracted until `--cgp-train-num-exons` exons are included. In practice, for vertebrate genomes, a few thousand exons corresponding to a few megabases of sequence are sufficient. If your genomes are more dense, this may vary. The trained model will be written to the AugustusCGP working directory, and can be used again on alignments with similar genomes.

AugustusPB

AugustusPB is a parameterization of AUGUSTUS to try and predict alternative isoforms using long range data. If any IsoSeq data are provided in the config file, and the --augustus-pb flag is set, the genomes with IsoSeq data will be run through and the results incorporated in the final gene set. AugustusPB runs on single whole genomes.

Modules

While the primary mode of operation of the pipeline is to launch the RunCat module, you may want to run specific modules. Any submodule can be ran by changing the luigi invocation to specify a submodule class instead.

PrepareFiles

This module parses the GFF3 annotation input, creating a genePred format file as well as a sqlite database. In addition, sequence files for all target genomes are extracted and converted to 2bit.

This module will populate the folders --work-dir/reference and --work-dir/genome_files.

Chaining

This step is the first precursor step to transMap. Pairwise genomic Kent-style chains are produced for each target genome from the designated reference. This step uses Toil and can be parallelized on a cluster.

This module will populate the folder --work-dir/chaining.

TransMap

This step runs transMap. The chain files are used to project annotations present in the GFF3 from the reference genome to each target genome.

EvaluateTransMap

This step performs the preliminary classification of transMap transcripts. This step populates the TransMapEvaluation table in the sqlite database for each target genome with the following classifiers:

1. AlnExtendsOffConfig: Does this alignment run off the end of a contig?
2. AlignmentPartialMap: Did this transcript not map completely?
3. AlnAbutsUnknownBases: Does this alignment have Ns immediately touching any exons?
4. PercentN: Percent of N bases in the alignment.

5. TransMapCoverage.
6. TransMapIdentity.
7. TransMapGoodness: A measure of alignment quality that takes into account both coverage and alignment size in the target. Related to Jim Kent's badness score.
8. TransMapOriginalIntronsPercent: The number of transMap introns within a wiggle distance of a intron in the parent transcript in transcript coordinates.
9. Synteny: Counts the number of genes in linear order that match up to +/- 5 genes.
10. ValidStart -- start with ATG?
11. ValidStop -- valid stop codon (in frame)?
12. ProperOrf -- is the orf a multiple of 3?

This module will populate the folder `--work-dir/transMap`.

FilterTransMap

This module runs the program `PslCDnaFilter` in a few different parameterizations. First, it runs the program with `-globalNearBest=0 -minSpan=0.2 -minCover=0.1`, which filters out alignments with coverage <10% and whose target genome span is less than 20% of the largest alignment, then uses the alignment score to pick the best single alignment. The output of this process is then split-gene resolved. This process involves identifying source genes that map to multiple contigs or disjoint on a single contig, identifying the highest scoring locus, then selecting alignments for transcripts in that locus that were filtered out by `globalNearBest`. The result is a single discrete locus for each source transcript. The plot `transmap_filtering.pdf` in the plots output gets populated by this, reporting on the number of alignments removed by span distance, coverage, and paralogy.

This module also runs `PslCDnaFilter` again, this time to discover paralogs. In this run, instead of `-globalNearBest` being set, `-localNearBest` is set to the user determined value (default is 0.15). This algorithm allows for multiple alignments from a given source to pass through if they are on different parts of the source transcript. In practice, this generally means filtering out truly paralogous alignments while retaining alignments of transcripts split across contigs. The alignments that are filtered out by this process are used to populate the `Paralogy` field in the final output, as well as the diagnostic plots.

The final step of this module is to resolve split genes. If transcripts for a given gene end up on different contigs or on disjoint intervals on the same contig, the locus with the highest average score is chosen and lower scoring alignments for the transcripts present in the other locus are rescued, if possible. This result populates the `Filter Mode = Rescued` part of transmap_filtering.pdf.

This module will further populate the folder `--work-dir/transMap`.

Augustus

As discussed above, this module runs `AugustusTM(R)`. If the pipeline is ran without a hints database, only the `AugustusTM` mode will be executed. This process is one of the most

computationally intensive steps, and should not be ran without a cluster.

This module will populate the folder `--work-dir/augustus`.

AugustusCgp

Running `AugustusCGP` is trickier than other modes. If your genomes are not closely related to an existing training set, you may need to perform logistic regression to train `AugustusCGP` before execution. A default parameter set is provided. This mode is also computationally intensive, and requires a cluster.

Each output transcript are assigned a parental gene, if possible. Parental gene assignment is done by looking to see if this transcript has at least `1` exonic base overlap with any filtered TransMap as well as unfiltered transMap. If the transcript overlaps more than one gene, the Jaccard metric (http://bedtools.readthedocs.io/en/latest/content/tools/jaccard.html) is used to try and resolve the ambiguity. If no gene stands out, this transcript is discarded. A sqlite table will record both the filtered and unfiltered overlaps.

Transcripts which are not assigned a parental gene will be considered ***novel*** in the consensus finding step. Most often, these are the result of gene family expansion or contraction in the reference. Looking at the raw `transMap` track in the final assembly hub will help resolve this.

This module will populate the folder `--work-dir/augustus_cgp`.

AugustusPb

Running `AugustusPB` requires that IsoSeq data be provided. This mode runs on single genomes, and attempts to discover new isoforms. Transcripts predicted in this process undergo the same parental gene assignment described above.

This module will populate the folder `--work-dir/augustus_pb`.

Hgm

`homGeneMapping` is a companion tool of `AugustusCGP`. This tool uses a HAL alignment to project RNA-seq and annotation information to target genomes. This is used to validate a splice junction in a target genome as being supported in one or more alternative genomes, as well as being supported in the reference annotation. This module populates the `*_Hgm` database table, where `*` is one of `transMap`, `augTM`, `augTMR`, `augCGP` or `augPB` depending on the transcripts being evaluated. This table has the following comma separated columns:

1. AllSpeciesIntronRnaSupport. The number of species with RNA-seq data supporting the intron junctions, in genomic order.
2. AllSpeciesExonRnaSupport. The number of species with RNA-seq data suporting the exons, in genomic order.
3. IntronRnaSupport. Same as #1, but only within this species.
4. ExonRnaSupport. Same as #2, but only within this species.

5. IntronAnnotSupport. A bit vector indicating if the intron junctions are supported by the reference annotation.
6. CdsAnnotSupport. A bit vector indicating if the CDS features are supported by the reference annotation.
7. ExonAnnotSupport. A bit vector indicating if the exon features are supported by the reference annotation.

This module will populate the folder `--work-dir/hgm`.

The output of the `homGeneMapping` module has more information embedded in the output files. Each GTF format file in the above folder has a added column on the end with a string like:

```
"0E-6273,1E-1524,2N:M*-1,3E-742,4E-1912,5E-1208"
```

Which can be interpreted as 'species 0 had 6273 extrinsic hints (RNA-seq coverage), species 1 has 1524 extrinsic hints, species 2 (the reference) had both a non-coding (N) and coding (M) junction', and so on. The species numeric values are at the top of the file, and correlate to the species ID assigned internally in the hints database. These data can be useful if you want to dig in to a specific annotation.

AlignTranscripts

Transcript alignment allows for `AugustusTM(R)` transcripts to be compared to their parental `transMap`. As a result, only protein coding transcripts are aligned. For each transcripts, alignment is performed by BLAT two ways -- in frame codon aware alignment, and mRNA alignment. The results of these alignments are saved in the folder `--work-dir/transcript_alignment`. These alignments are used to create functional annotations of transcripts in the EvaluateTranscripts module.

EvaluateTranscripts

A series of classifiers that evaluate transcript pairwise alignments for `transMap` and `AugustusTM(R)` output.

These classifiers are broken down into 2 groups, which will each end up as a table in the database:

__Metrics:

These classifiers are per-transcript evaluations based on both the transcript alignment and the genome context.

1. PercentUnknownBases: % of mRNA bases that are Ns.
2. AlnCoverage: Alignment coverage in transcript space.
3. AlnIdentity: Alignment identity in transcript space.
4. OriginalIntrons. Original introns is a bit vector that evaluates whether the intron junctions in transcript coordinate space are within 5 bases either direction from the original transcript. This is a powerful approach to identifying retroposed pseudogenes

or problematic alignments.

5. ValidStart -- start with ATG?
6. ValidStop -- valid stop codon (in frame)?
7. ProperOrf -- is the orf a multiple of 3?

__Evaluation:

These classifiers are per-transcript evaluations based on the transcript alignment. Unlike the other two tables, this table stores the actual location of the problems (in genome coordinates) as a BED-like format. In cases where there are multiple problems, they will be additional rows.

1. CodingInsertion: Do we have any frame-shifting coding insertions?
2. CodingDeletion: Do we have any frame-shifting coding deletions?
3. CodingMult3Insertion: Do we have any mod3 coding insertions?
4. CodingMult3Deletion: Do we have any mod3 coding deletions?
5. NonCodingInsertion: Do we have indels in UTR sequence?
6. NonCodingDeletion: Do we have any indels in UTR sequence?
7. InFrameStop: Are there any in-frame stop codons?

Where txMode is one of transMap, augTM, augTMR and alnMode is one of CDS or mRNA.

The evaluation tables will be loaded as tracks in the final assembly hub.

Consensus

The consensus finding process takes in transcripts from every mode and attempts to find the highest quality ortholog for a source transcript. The de-novo transcript modes are also evaluated for providing novel isoforms or novel loci. The final gene set is output with a series of features measuring how confident the prediction is.

To evaluate `transMap`, `AugustusTM` and `AugustusTMR` transcripts a consensus score is assigned to each. This score is the sum of the alignment goodness, intron/exon annotation support, original intron support, and intron/exon RNA-seq/IsoSeq support if extrinsic data were provided. The transcript with the highest consensus score is chosen.

If one of the de-novo `augustus` modes is run, then the those transcripts are evaluated for providing novel information. If a prediction did not overlap any transMap projections, then it is tagged as putative novel and incorporated into the gene set. If a prediction overlaps a `transMap` projection that was filtered out during paralog resolution, then it is tagged as a possible paralog as well as with the names of overlapping transcripts and incorporated into the gene set. If a prediction overlaps a transMap projection and contains a splice junction not seen in the reference annotation, then it is tagged as a novel isoform and incorporated into the gene set as a member of the gene it overlapped with.

After consensus finding is complete, a final filtering process is performed. This filtering process deduplicates the transcript set. Duplicates most often occur when the `augustus` execution modes create an identical transcript model from different input isoforms. In this case, the duplicates are removed and the remaining transcript tagged with the names of

alternative source transcripts. Finally, strand resolution throws out transcripts that are on opposite strands. The correct strand is chosen by looking at which contains the most high quality transcripts.

After consensus finding, a final output gene set is produced in both `GFF3` and `genePred` format. The `genePred` annotations also have a additional `.gp_info` file that has the additional fields described below.

The output will appear in `--output-dir/consensus`.

Plots

A large range of plots are produced in `--output-dir/plots`. These include:

1. `denovo.pdf`. If either ***de-novo*** mode was ran, this plot will report the results. See the above description of the tags in consensus or GFF3 tags sections.
2. `completeness.pdf`: The number of genes/transcripts successfully mapped over. The top of the x axis is marked with a red line representing the amount of genes/transcripts the source annotation had.
3. `consensus_extrinsic_support`: A violin plot of the level of extrinsic support seen across all species for splices and exons, as found by `homGeneMapping`. Provides a overall plot and a per-biotype plot.
4. `consensus_anotation_support`: A violin plot of the level of annotation support seen across all species for splices and exons, as found by `homGeneMapping`. Provides a overall plot and a per-biotype plot.
5. `coverage.pdf`: A violinplot that shows the overall transcript coverage in the ***consensus*** set. Provides a overall plot and a per-biotype plot.
6. `identity.pdf`: A violinplot that shows the overall transcript identity in the ***consensus*** set. Provides a overall plot and a per-biotype plot.
7. `transmap_coverage.pdf`: A violinplot that shows the overall transcript coverage in the filtered transMap output. Provides a overall plot and a per-biotype plot.
8. `transmap_ identity.pdf`: A violinplot that shows the overall transcript identity in the filtered transMap output. Provides a overall plot and a per-biotype plot.
9. `missing_genes-transcripts.pdf`: Similar to `completeness.pdf`, this plot reports the number of genes and transcripts in the original annotation set not found on the target genomes.
10. `paralogy.pdf`: Stacked bar charts of the number of alignments a given source transcript had in each target.
11. `split_genes.pdf`: The number of genes split within and between contigs.
12. `transcript_modes.pdf`: The number of modes that supported a given comparative annotation. Applies only to protein coding transcripts derived from `transMap`, because `AugustusTMR` is not ran on non-coding inputs.
13. `augustus_improvement.pdf`: A scatterplot + density plot reporting the improvement of primary consensus metrics when an `augustus` transcript was chosen over a transMap transcript. The density plot may fail in some cases.

14. `coding_indels.pdf`: The rate of insertions, deletions and indels that are a multiple of 3 are reported from the final consensus set based on the pairwise alignments. Preference is given to the CDS space alignment, if it worked.
15. `IsoSeq_isoform_validation.pdf`: The number of transcripts in the consensus set whose intron structure is exactly validated by at least one IsoSeq read.

GFF3 tags:

1. `gene_id`: Unique gene ID assigned to this gene.
2. `transcript_id`: Unique transcript ID assigned to this gene.
3. `alignment_id`: Original alignment ID internally used in the pipeline. Provides a link to the gene sets input to consensus finding.
4. `alternative_source_transcripts`: If deduplication collapsed transcripts, report the other `source_transcript` IDs.
5. `exon_annotation_support`: Was this exon supported by the reference annotation?
6. `exon_rna_support`: Was this exon supported by the extrinsic database?
7. `frameshift`: Is this transcript frameshifted relative to `source_transcript`?
8. `gene_biotype`: The `source_gene` biotype. If this is a ***de-novo*** prediction, this field will say unknown_likely_coding.
9. `intron_annotation_support`: Was this intron supported by the reference annotation?
10. `intron_rna_support`: Was this intron supported by the extrinsic database?
11. `source_gene`: The gene ID of the source gene, if this is a projection transcript.
12. `source_gene_common_name`: The common name of the source gene.
13. `source_transcript`: The ID of the source transcript.
14. `transcript_biotype`: The biotype of the source transcript, or unknown_likely_coding for ***de-novo*** predictions.
15. `transcript_class`: For projection transcripts, just says ortholog. For ***de-novo*** transcripts, will be one of poor_alignment, possible_paralog, putative_novel_isoform, or putative_novel. See the consensus finding section for descriptions.
16. `transcript_modes`: Comma separated list of transcript modes. The same information as the transcript_modes.pdf plot.
17. `pacbio_isoform_supported`: Was this isoform supported by at least one IsoSeq read?
18. `paralogy`: Comma separated list of alignments identified as possible paralogs for this transcript.
19. `gene_alternate_contigs`: If this gene was split across multiple contigs, this will have a comma separated list of alternative locations.

For `GFF3` output, the alignment goodness is in the score field. For `.gp_info`, it is a column. For `.gp_info`, the support features are collapsed into comma separated vectors instead of being on their respective features.

Bibliography

[1] Saul B. Needleman and Christian D. Wunsch. "A general method applicable to the search for similarities in the amino acid sequence of two proteins". In: *Journal of Molecular Biology* 48.3 (1970), pp. 443–453. ISSN: 0022-2836. DOI: http://dx.doi.org/10.1016/0022-2836(70)90057-4. URL: http://www.sciencedirect.com/science/article/pii/0022283670900574.

[2] T. Smith and M. Waterman. "Identification of common molecular subsequences". In: *Journal of Molecular Biology* 147.1 (1981), pp. 195–197. ISSN: 00222836. DOI: 10.1016/0022-2836(81)90087-5. URL: http://dx.doi.org/10.1016/0022-2836(81)90087-5.

[3] Stephen Altschul et al. "Basic local alignment search tool". In: *J. Mol. Biol.* 215 (1990), pp. 403–410.

[4] S Batzoglou. "The many faces of sequence alignment". In: *Brief Bioinform* 6.1 (2005), pp. 6–22. ISSN: 1467-5463. DOI: 10.1093/bib/6.1.6.

[5] Nick Bray, Inna Dubchak, and Lior Pachter. "AVID: A global alignment program." In: *Genome research* 13.1 (2003), pp. 97–102. ISSN: 10889051. DOI: 10.1101/gr.789803.

[6] Arthur L. Delcher et al. "Alignment of whole genomes". In: *Nucleic Acids Research* 27.11 (1999), pp. 2369–2376. ISSN: 03051048. DOI: 10.1093/nar/27.11.2369.

[7] Michael Brudno et al. "LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA". In: *Genome Research* 13.4 (2003), pp. 721–731. ISSN: 10889051. DOI: 10.1101/gr.926603.

[8] W James Kent et al. “Evolution’s cauldron: duplication, deletion, and rearrangement in the mouse and human genomes.” In: *Proceedings of the National Academy of Sciences of the United States of America* 100.20 (2003), pp. 11484–11489. ISSN: 0027-8424. DOI: 10.1073/pnas.1932072100.

[9] Scott Schwartz et al. “Human–mouse alignments with BLASTZ”. In: *Genome research* 13.1 (2003), pp. 103–107.

[10] Mathieu Blanchette et al. “Aligning multiple genomic sequences with the threaded blockset aligner”. In: *Genome Research* 14.4 (2004), pp. 708–715. ISSN: 10889051. DOI: 10.1101/gr.1933104.

[11] Toby Johnson. “Reciprocal best hits are not a logically sufficient condition for orthology”. In: *arXiv* 5.2 (2007), pp. 1–8. arXiv: 0706.0117. URL: http://arxiv.org/abs/0706.0117.

[12] Eugene V. Koonin. “Orthologs, Paralogs, and Evolutionary Genomics”. In: *Annual Review of Genetics* 39.1 (2005), pp. 309–338. ISSN: 0066-4197. DOI: 10.1146/annurev.genet.39.073003.114725.

[13] T Jiang and L Wang. “On the complexity of multiple sequence alignment”. In: *Journal of Computational Biology* 1 (1994), pp. 337–348.

[14] Da-Fei Feng and Russell F Doolittle. “Progressive sequence alignment as a prerequisite to correct phylogenetic trees”. In: *Journal of molecular evolution* 25.4 (1987), pp. 351–360.

[15] Benedict Paten et al. “Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs.” In: *Genome research* 18.11 (2008), pp. 1814–28. ISSN: 1088-9051. DOI: 10.1101/gr.076554.108. URL: http://genome.cshlp.org.ezp-prod1.hul.harvard.edu/content/18/11/1814.

[16] Benjamin Raphael et al. “A novel method for multiple alignment of sequences with repeated and shuffled elements”. In: *Genome Research* 14.11 (2004), pp. 2336–2346. ISSN: 10889051. DOI: 10.1101/gr.2657504.

[17] Bronwen L. Aken et al. “Ensembl 2017”. In: *Nucleic Acids Research* 45.D1 (2017), pp. D635–D642. ISSN: 13624962. DOI: 10.1093/nar/gkw1104.

[18] G. Hickey et al. “HAL: a hierarchical format for storing and analyzing multiple genome alignments”. In: *Bioinformatics* 29.10 (2013), pp. 1341–1342. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btt128. URL: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btt128.

[19] Ngan Nguyen et al. “Comparative assembly hubs: Web-accessible browsers for comparative genomics”. In: *Bioinformatics* 30.23 (2014), pp. 3293–3301. ISSN: 14602059. DOI: 10.1093/bioinformatics/btu534. arXiv: arXiv:1311.1241v1.

[20] R.S. Harris. “Improved pairwise alignment of genomic DNA”. PhD thesis. The Pennsylvania State University, 2007.

[21] Bin Ma, John Tromp, and Ming Li. “PatternHunter: faster and more sensitive homology search.” In: *Bioinformatics (Oxford, England)* 18.3 (2002), pp. 440–445. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/18.3.440.

[22] Szymon M Kiełbasa et al. “Adaptive seeds tame genomic sequence comparison”. In: *Genome research* 21.3 (2011), pp. 487–493.

[23] Jonathan Casper et al. “The UCSC Genome Browser database: 2018 update”. In: *Nucleic Acids Research* December 2017 (2017), pp. 1–8. ISSN: 0305-1048. DOI: 10.1093/nar/gkx985. URL: http://fdslive.oup.com/www.oup.com/pdf/production%7B%5C_%7Din%7B%5C_%7Dprogress.pdf.

[24] Benedict Paten et al. “Genome-wide nucleotide-level mammalian ancestor reconstruction”. In: *Genome Research* 18.11 (2008), pp. 1829–1843. ISSN: 10889051. DOI: 10.1101/gr.076521.108.

[25] Bronwen L. Aken et al. “The Ensembl gene annotation system”. In: *Database* 2016 (2016). DOI: 10.1093/database/baw093. eprint: http://database.oxfordjournals.org/content/2016/baw093.full.pdf+html. URL: http://database.oxfordjournals.org/content/2016/baw093.abstract.

[26] B. Paten et al. “Cactus graphs for genome comparisons”. In: *J. Comput. Biol.* 18.3 (Mar. 2011), pp. 469–481.

[27] B. Paten et al. “Cactus: Algorithms for genome multiple sequence alignment”. In: *Genome Res.* 21.9 (Sept. 2011), pp. 1512–1528.

[28] Dent Earl et al. “Alignathon : a competitive assessment of whole-genome alignment methods”. In: *Genome research* (2014), pp. 2077–2089. DOI: 10.1101/gr.174920.114.Freely.

[29] Ngan Nguyen et al. “Building a pan-genome reference for a population.” In: *Journal of computational biology : a journal of computational molecular cell biology* 22.5 (2015), pp. 387–401. ISSN: 1557-8666. DOI: 10.1089/cmb.2014.0146. URL: http://online.liebertpub.com/doi/10.1089/cmb.2014.0146.

[30] Jennifer Harrow et al. “GENCODE: the reference human genome annotation for The ENCODE Project”. In: *Genome research* 22.9 (2012), pp. 1760–1774.

[31] Jiang Zhu et al. “Genome-wide chromatin state transitions associated with developmental and environmental cues”. In: *Cell* 152.3 (2013), pp. 642–654.

[32] ENCODE Project Consortium et al. “The ENCODE (ENCyclopedia of DNA elements) project”. In: *Science* 306.5696 (2004), pp. 636–640.

[33] Aimée M Deaton and Adrian Bird. “CpG islands and the regulation of transcription”. In: *Genes & development* 25.10 (2011), pp. 1010–1022.

[34] Stephen T Sherry et al. “dbSNP: the NCBI database of genetic variation”. In: *Nucleic acids research* 29.1 (2001), pp. 308–311.

[35] Stanley I Letovsky et al. “GDB: the human genome database”. In: *Nucleic Acids Research* 26.1 (1998), pp. 94–99.

[36] Alexander V Lukashin and Mark Borodovsky. “GeneMark. hmm: new solutions for gene finding”. In: *Nucleic acids research* 26.4 (1998), pp. 1107–1115.

[37] David Kulp David Haussler and Martin G Reese Frank H Eeckman. “A generalized hidden Markov model for the recognition of human genes in DNA”. In: *Proc. Int. Conf. on Intelligent Systems for Molecular Biology, St. Louis.* 1996, pp. 134–142.

[38] Kim D Pruitt, Tatiana Tatusova, and Donna R Maglott. “NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins”. In: *Nucleic acids research* 35.suppl_1 (2006), pp. D61–D65.

[39] Brandi L Cantarel et al. "MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes". In: *Genome research* 18.1 (2008), pp. 188–196.

[40] David Gordon et al. "Long-read sequence assembly of the gorilla genome". In: *Science* 352.6281 (2016), aae0344.

[41] *CHROMIUM de novo Assembly Solution*. Tech. rep. URL: http://go.10xgenomics.com/l/172142/2016-08-10/3svkc/172142/8088/LIT00005_RevA_Chromium_De_Novo_Assembly_Solution_Application_Note2.pdf.

[42] David Haussler et al. "Genome 10K: a proposal to obtain whole-genome sequence for 10 000 vertebrate species". In: *Journal of Heredity* 100.6 (2009), pp. 659–674.

[43] Gene E Robinson et al. "Creating a buzz about insect genomes". In: *Science* 331.6023 (2011), pp. 1386–1386.

[44] Robert H Waterston and Lior Pachter. "Initial sequencing and comparative analysis of the mouse genome". In: *Nature* 420.6915 (2002), pp. 520–562.

[45] Paul Flicek et al. "Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map". In: *Genome research* 13.1 (2003), pp. 46–54.

[46] Thomas Wiehe et al. "SGP-1: prediction and validation of homologous genes based on sequence alignments". In: *Genome Research* 11.9 (2001), pp. 1574–1583.

[47] Marina Alexandersson, Simon Cawley, and Lior Pachter. "SLAM: cross-species gene finding and alignment with a generalized pair hidden Markov model". In: *Genome Research* 13.3 (2003), pp. 496–502.

[48] Ru-Fang Yeh, Lee P Lim, and Christopher B Burge. "Computational inference of homologous gene structures in the human genome". In: *Genome research* 11.5 (2001), pp. 803–816.

[49] Mikhail S Gelfand, Andrey A Mironov, and Pavel A Pevzner. "Gene recognition via spliced sequence alignment". In: *Proceedings of the National Academy of Sciences* 93.17 (1996), pp. 9061–9066.

[50] Samuel S Gross and Michael R Brent. "Using multiple alignments to improve gene prediction". In: *Journal of computational biology* 13.2 (2006), pp. 379–393.

[51] Marijke J van Baren, Brian C Koebbe, and Michael R Brent. "Using N-SCAN or TWINSCAN to Predict Gene Structures in Genomic DNA Sequences". In: *Current Protocols in Bioinformatics* (2007), pp. 4–8.

[52] Paul Flicek. "Gene prediction: compare and CONTRAST". In: *Genome biology* 8.12 (2007), p. 233.

[53] Samuel S Gross et al. "CONTRAST: a discriminative, phylogeny-free approach to multiple informant de novo gene prediction". In: *Genome biology* 8.12 (2007), R269.

[54] John Lafferty, Andrew McCallum, and Fernando CN Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data". In: 2001.

[55] Mario Stanke et al. "Using native and syntenically mapped cDNA alignments to improve de novo gene finding". In: *Bioinformatics* 24.5 (2008), pp. 637–644.

[56] KJ Hoff and M Stanke. "Current methods for automated annotation of protein-coding genes". In: *Current Opinion in Insect Science* 7 (2015), pp. 8–14.

[57] Mammalian Gene Collection (MGC) Program Team et al. "Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences". In: *Proceedings of the National Academy of Sciences* 99.26 (2002), pp. 16899–16903.

[58] Dennis A Benson et al. "GenBank". In: *Nucleic acids research* 28.1 (2000), pp. 15–18.

[59] Chaochun Wei and Michael R Brent. "Using ESTs to improve the accuracy of de novo gene prediction". In: *BMC bioinformatics* 7.1 (2006), p. 327.

[60] Ewan Birney, Michele Clamp, and Richard Durbin. "GeneWise and genomewise". In: *Genome research* 14.5 (2004), pp. 988–995.

[61] Mario Stanke et al. "Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources". In: *BMC bioinformatics* 7.1 (2006), p. 62.

[62] Mario Stanke et al. “AUGUSTUS: a web server for gene finding in eukaryotes”. In: *Nucleic acids research* 32.suppl 2 (2004), W309–W312.

[63] Mark Yandell and Daniel Ence. “A beginner’s guide to eukaryotic genome annotation”. In: *Nature Reviews Genetics* 13.5 (2012), pp. 329–342.

[64] Evgeni M Zdobnov and Rolf Apweiler. “InterProScan–an integration platform for the signature-recognition methods in InterPro”. In: *Bioinformatics* 17.9 (2001), pp. 847–848.

[65] Oliver Keller et al. “A novel hybrid gene prediction method employing protein multiple sequence alignments”. In: *Bioinformatics* 27.6 (2011), pp. 757–763.

[66] GTEx Consortium et al. “The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans”. In: *Science* 348.6235 (2015), pp. 648–660.

[67] Xinxia Peng et al. “Integrative deep sequencing of the mouse lung transcriptome reveals differential expression of diverse classes of small RNAs in response to respiratory virus infection”. In: *MBio* 2.6 (2011), e00198–11.

[68] Ian T Fiddes et al. “Comparative Annotation Toolkit (CAT)-simultaneous clade and personal genome annotation”. In: *bioRxiv* (2017), p. 231118.

[69] Virag Sharma, Anas Elghafari, and Michael Hiller. “Coding exon-structure aware realigner (CESAR) utilizes genome alignments for accurate comparative gene annotation”. In: *Nucleic acids research* 44.11 (2016), e103–e103.

[70] Irmtraud M Meyer and Richard Durbin. “Gene structure conservation aids similarity based gene prediction”. In: *Nucleic acids research* 32.2 (2004), pp. 776–783.

[71] Liliana Florea et al. “Gene and alternative splicing annotation with AIR”. In: *Genome research* 15.1 (2005), pp. 54–66.

[72] Stefanie König et al. “Simultaneous gene finding in multiple genomes”. In: *Bioinformatics* 32.21 (2016).

[73] Alona Sosinsky, Gustavo Glusman, and Doron Lancet. “The genomic structure of human olfactory receptor genes”. In: *Genomics* 70.1 (2000), pp. 49–61.

[74] Albert J Vilella et al. "EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates". In: *Genome research* 19.2 (2009), pp. 327–335.

[75] Elena Rivas and Sean R Eddy. "Noncoding RNA gene detection using comparative sequence analysis". In: *BMC bioinformatics* 2.1 (2001), p. 8.

[76] Julien Lagarde et al. "High-throughput annotation of full-length long noncoding RNAs with Capture Long-Read Sequencing (CLS)". In: *bioRxiv* (2017). DOI: 10.1101/105064. eprint: http://biorxiv.org/content/early/2017/02/01/105064.full.pdf. URL: http://biorxiv.org/content/early/2017/02/01/105064.

[77] Sven Diederichs. "The four dimensions of noncoding RNA conservation". In: *Trends in genetics* 30.4 (2014), pp. 121–123.

[78] Igor Ulitsky and David P Bartel. "lincRNAs: genomics, evolution, and mechanisms". In: *Cell* 154.1 (2013), pp. 26–46.

[79] Michael F Lin, Irwin Jungreis, and Manolis Kellis. "PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions". In: *Bioinformatics* 27.13 (2011), pp. i275–i282.

[80] Chen-Shan Chin et al. "Phased diploid genome assembly with single-molecule real-time sequencing". In: *Nature Methods* 13.12 (2016), pp. 1050–1054.

[81] Jonas Korlach et al. "De Novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads". In: *GigaScience* (2017), gix085.

[82] Bruce J Walker et al. "Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement". In: *PloS one* 9.11 (2014), e112963.

[83] Erik Garrison and Gabor Marth. "Haplotype-based variant detection from short-read sequencing". In: *arXiv preprint arXiv:1207.3907* (2012).

[84] Noah Spies et al. "Genome-wide reconstruction of complex structural variants using read clouds". In: *bioRxiv* (2016), p. 074518.

[85] Jacob O Kitzman. "Haplotypes drop by drop: short-read sequencing provides haplotype information when long DNA fragments are barcoded in microfluidic droplets". In: *Nature biotechnology* 34.3 (2016), pp. 296–299.

[86] Alex Bishara et al. "Read clouds uncover variation in complex regions of the human genome". In: *Genome research* 25.10 (2015), pp. 1570–1580.

[87] Serafim Batzoglou et al. "Human and mouse gene structure: comparative analysis and application to exon prediction". In: *Genome research* 10.7 (2000), pp. 950–958.

[88] Jakob Skou Pedersen and Jotun Hein. "Gene finding with a hidden Markov model of genome structure and evolution". In: *Bioinformatics* 19.2 (2003), pp. 219–227.

[89] Adam Siepel and David Haussler. "Computational identification of evolutionarily conserved exons". In: *Proceedings of the eighth annual international conference on Resaerch in computational molecular biology*. ACM. 2004, pp. 177–186.

[90] David Carter and Richard Durbin. "Vertebrate gene finding from multiple-species alignments using a two-level strategy". In: *Genome Biology* 7.1 (2006), S6.

[91] Brian J Haas et al. "Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments". In: *Genome biology* 9.1 (2008), R7.

[92] Brian J Haas et al. "De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity". In: *Nature protocols* 8.8 (2013).

[93] Brian J Haas et al. "Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies". In: *Nucleic acids research* 31.19 (2003), pp. 5654–5666.

[94] Michael Brudno et al. "Glocal alignment: Finding rearrangements during alignment". In: *Bioinformatics* 19.SUPPL. 1 (2003). ISSN: 13674803. DOI: 10.1093/bioinformatics/btg1005.

[95] Inna Dubchak et al. "Multiple whole-genome alignments without a reference organism". In: *Genome Research* 19.4 (2009), pp. 682–689. ISSN: 10889051. DOI: 10.1101/gr.081778.108.

[96] Aaron E. Darling, Bob Mau, and Nicole T. Perna. "Progressivemauve: Multiple genome alignment with gene gain, loss and rearrangement". In: *PLoS ONE* 5.6 (2010). ISSN: 19326203. DOI: 10.1371/journal.pone.0011147.

[97] Dent Earl. "EVALUATION AND APPLICATION OF THE THREE AŚ OF GENOMICS: ASSEMBLY, ALIGNMENT, ANNOTATION." PhD thesis. UC Santa Cruz, 2014.

[98] W James Kent et al. "Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes". In: *Proceedings of the National Academy of Sciences* 100.20 (2003), pp. 11484–11489.

[99] Jingchun Zhu et al. "Comparative genomics search for losses of long-established genes on the human lineage". In: *PLoS computational biology* 3.12 (2007), e247.

[100] W James Kent. "BLAT the BLAST like alignment tool". In: *Genome research* 12.4 (2002), pp. 656–664.

[101] Ngan Nguyen et al. "Comparative assembly hubs: web-accessible browsers for comparative genomics". In: *Bioinformatics* 30.23 (2014), pp. 3293–3301.

[102] Kenneth Paigen. "One hundred years of mouse genetics: an intellectual history. II. The molecular revolution (1981-2002)." In: *Genetics* 163.4 (Apr. 2003), pp. 1227–35.

[103] Jackson Laboratory. *Early history of mouse genetics*. URL: http://research.jax.org/mousegenetics/development/history.html.

[104] Atsushi Yoshiki and Kazuo Moriwaki. "Mouse Phenome Research: Implications of Genetic Background". In: *ILAR Journal* 47.2 (2006), pp. 94–102.

[105] Jean Louis Guénet. "The mouse genome." In: *Genome Research* 15.12 (Dec. 2005), pp. 1729–40.

[106] The Knockout Mouse Consortium. "The Knockout Mouse Project". In: *Nature Genetics* 36 (2004).

[107] J. A. Beck et al. "Genealogies of mouse inbred strains." In: *Nature Genetics* 24.1 (Jan. 2000), pp. 23–5.

[108] Kevin K. Ohlemiller. “Contributions of mouse models to understanding of age- and noise-related hearing loss.” In: *Brain Research* 1091.1 (May 2006), pp. 89–102.

[109] J. L. Peirce et al. “A major influence of sex-specific loci on alcohol preference in C57Bl/6 and DBA/2 inbred mice.” In: *Mammalian Genome* 9.12 (Dec. 1998), pp. 942–8.

[110] Edward Leiter, Michal Prochazka, and Douglas Coleman. “The Non-Obese Diabetic (NOD) Mouse”. In: *American Journal of Pathology* 128 (1987).

[111] Bradford D. Haskell et al. “The diabetes-prone NZO/HlLt strain. I. Immunophenotypic comparison to the related NZB/BlNJ and NZW/LacJ strains.” In: *Laboratory investigation; a journal of technical methods and pathology* 82.7 (July 2002), pp. 833–42.

[112] Jonathan Flint and Trudy F. C. Mackay. “Genetic architecture of quantitative traits in mice, flies, and humans.” In: *Genome research* 19.5 (May 2009), pp. 723–33.

[113] William Valdar et al. “Genome-wide genetic association of complex traits in heterogeneous stock mice.” In: *Nature genetics* 38.8 (Aug. 2006), pp. 879–87.

[114] The Mouse Genomes Project. *The Mouse Genomes Project*. URL: `http://www.sanger.ac.uk/resources/mouse/genomes/`.

[115] Molly A. Bogue et al. “Mouse Phenome Database (MPD)”. In: *Nucleic Acids Research* 35.1 (Jan. 2007), pp. D643–9.

[116] Kelly A. Frazer et al. “A sequence-based variation map of 8.27 million SNPs in inbred mouse strains.” In: *Nature* 448.7157 (Aug. 2007), pp. 1050–3.

[117] Thomas M. Keane et al. “Mouse genomic variation and its effect on phenotypes and gene regulation”. In: *Nature* 477.289 (2011).

[118] Jared Simpson and Richard Durbin. “Efficient de novo assembly of large genomes using compressed data structures”. In: *Genome Research* (2012).

[119] P. Latreille et al. “Optical mapping as a routine tool for bacterial genome sequence finishing”. In: *BMC genomics* 8.1471-2164 (2007), p. 321.

[120] Mikhail Kolomogorov et al. “Ragout—a reference-assisted assembly tool for bacterial genomes”. In: *Bioinformatics* 30 (2014).

[121] Benedict Paten et al. “Cactus: Algorithms for genome multiple sequence alignment”. In: *Genome Research* (2011).

[122] D. W. Bailey. “Sources of subline divergence and their relative importance for sublines of six major inbred strains of mice.” In: Origins of inbred mice. Academic Press, 1978.

[123] The ENCODE Project Consortium. “An integrated encyclopedia of DNA elements in the human genome”. In: *Nature* 489 (2012), pp. 57–74.

[124] Jennifer Harrow et al. “GENCODE: The reference human genome annotation for The ENCODE Project”. In: *Genome Research* 22 (2012), pp. 1760–1774.

[125] The 1000 Genomes Project Consortium. “A map of human genome variation from population-scale sequencing”. In: *Nature* 467 (2010).

[126] Susan Carpenter et al. “A long noncoding RNA mediates both activation and repression of immune response genes”. In: *science* 341.6147 (2013), pp. 789–792.

[127] Nicholas H Putnam et al. “Chromosome-scale shotgun assembly using an in vitro method for long-range linkage”. In: *Genome research* 26.3 (2016), pp. 342–350.

[128] Miten Jain et al. “Nanopore sequencing and assembly of a human genome with ultra-long reads”. In: *bioRxiv* (2017). DOI: 10.1101/128835. eprint: http://www.biorxiv.org/content/early/2017/04/20/128835.full.pdf. URL: http://www.biorxiv.org/content/early/2017/04/20/128835.

[129] Mark JP Chaisson, Richard K Wilson, and Evan E Eichler. “Genetic variation and the de novo assembly of human genomes”. In: *Nature Reviews Genetics* 16.11 (2015), pp. 627–640.

[130] David Thybert et al. “Repeat associated mechanisms of genome evolution and function revealed by the Mus caroli and Mus pahari genomes”. In: *bioRxiv* (2017). DOI: 10.1101/158659. eprint: http://www.biorxiv.org/content/early/2017/07/02/158659.full.pdf. URL: http://www.biorxiv.org/content/early/2017/07/02/158659.

[131] Erich D Jarvis et al. “Whole-genome analyses resolve early branches in the tree of life of modern birds”. In: *Science* 346.6215 (2014), pp. 1320–1331.

[132] Valerie A Schneider et al. “Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly”. In: *Genome research* 27.5 (2017), pp. 849–864.

[133] Karyn Meltz Steinberg et al. “Single haplotype assembly of the human genome from a hydatidiform mole”. In: *Genome research* 24.12 (2014), pp. 2066–2076.

[134] Matthew Pendleton et al. “Assembly and diploid architecture of an individual human genome via single-molecule technologies”. In: *Nature methods* 12.8 (2015), pp. 780–786.

[135] Monte Westerfield et al. “Zebrafish informatics and the ZFIN database”. In: *Methods in cell biology* 60 (1998), pp. 339–355.

[136] Lincoln Stein et al. “WormBase: network access to the genome and biology of Caenorhabditis elegans”. In: *Nucleic acids research* 29.1 (2001), pp. 82–86.

[137] David Swarbreck et al. “The Arabidopsis Information Resource (TAIR): gene structure and function annotation”. In: *Nucleic acids research* 36.suppl 1 (2008), pp. D1009–D1014.

[138] Benedict Paten et al. “Cactus: Algorithms for genome multiple sequence alignment”. In: *Genome research* 21.9 (2011), pp. 1512–1528.

[139] Mathieu Blanchette et al. “Aligning multiple genomic sequences with the threaded blockset aligner”. In: *Genome research* 14.4 (2004), pp. 708–715.

[140] Dent Earl et al. “Alignathon: a competitive assessment of whole-genome alignment methods”. In: *Genome research* 24.12 (2014), pp. 2077–2089.

[141] Webb Miller et al. “28-way vertebrate alignment and conservation track in the UCSC Genome Browser”. In: *Genome research* 17.12 (2007), pp. 1797–1808.

[142] Sean P Gordon et al. “Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing”. In: *PloS one* 10.7 (2015), e0132628.

[143] Glenn Hickey et al. “HAL: a hierarchical format for storing and analyzing multiple genome alignments”. In: *Bioinformatics* (2013), btt128.

[144] John Huddleston et al. “Discovery and genotyping of structural variation from long-read haploid genome sequence data”. In: *Genome Research* (2016), gr–214007.

[145] Neil I Weisenfeld et al. “Direct determination of diploid genome sequences”. In: *bioRxiv* (2016). DOI: 10.1101/070425. eprint: http://biorxiv.org/content/early/2016/08/19/070425.full.pdf. URL: http://biorxiv.org/content/early/2016/08/19/070425.

[146] Konrad J Karczewski et al. “The ExAC browser: displaying reference data information from over 60 000 exomes”. In: *Nucleic acids research* 45.D1 (2016), pp. D840–D845.

[147] Zhiyong Shi et al. “Association of TRB3 gene Q84R polymorphism with type 2 diabetes mellitus in Chinese population”. In: *Endocrine* 35.3 (2009), pp. 414–419.

[148] Donna Karolchik et al. “The UCSC genome browser database”. In: *Nucleic acids research* 31.1 (2003), pp. 51–54.

[149] Alexey A Fushan et al. “Gene expression defines natural changes in mammalian lifespan”. In: *Aging cell* 14.3 (2015), pp. 352–365.

[150] Diego Cortez et al. “Origins and functional evolution of Y chromosomes across mammals”. In: *Nature* 508.7497 (2014), pp. 488–493.

[151] Wei Liu et al. “Identification of distinct genes associated with seawater aspiration-induced acute lung injury by gene expression profile analysis”. In: *Molecular Medicine Reports* 14.4 (2016), pp. 3168–3178.

[152] Felipe A Simão et al. “BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs”. In: *Bioinformatics* 31.19 (2015), pp. 3210–3212.

[153] A Byrne et al. “Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface”. In: *Issue: NATURE COMMUNICATIONS, 8* (2017).

[154] John Vivian et al. “Toil enables reproducible, open source, big biomedical data analyses”. In: *Nature biotechnology* 35.4 (2017), p. 314.

[155] Alexander Dobin et al. “STAR: ultrafast universal RNA-seq aligner”. In: *Bioinformatics* 29.1 (2013), pp. 15–21.

[156] Nicolas Bray et al. “Near-optimal RNA-Seq quantification”. In: *arXiv preprint arXiv:1505.02710* (2015).

[157] Aaron R Quinlan and Ira M Hall. “BEDTools: a flexible suite of utilities for comparing genomic features”. In: *Bioinformatics* 26.6 (2010), pp. 841–842.

[158] Thomas D Wu and Colin K Watanabe. “GMAP: a genomic mapping and alignment program for mRNA and EST sequences”. In: *Bioinformatics* 21.9 (2005), pp. 1859–1875.