Annotating genes and genomes with DNA sequences extracted from biomedical articles

Maximilian Haussler
Martin Gerner
Casey Bergman

University of Manchester, UK

Which article is analyzing this genomic region?

Which genomic regions are analyzed in this article?

Human Molecular Genetics, 2008, Vol. 17, No. 23 **3740–3760**
doi:10.1093/hmg/ddn271
Advance Access published on September 16, 2008

Identification of *Arx* transcriptional targets in the developing basal forebrain

Carl T. Fulp[1], Ginam Cho[2], Eric D. Marsh[3], Ilya M. Nasrallah[2], Patricia A. Labosky[4] and Jeffrey A. Golden[1,2,*]

[1]Neuroscience Graduate Group, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA, [2]Department of Pathology and [3]Division of Neurology, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA and [4]Vanderbilt Center for Stem Cell Biology, Vanderbilt University Medical Center, Nashville, TN 37232, USA

Received June 4, 2008; Revised August 1, 2008; Accepted August 27, 2008

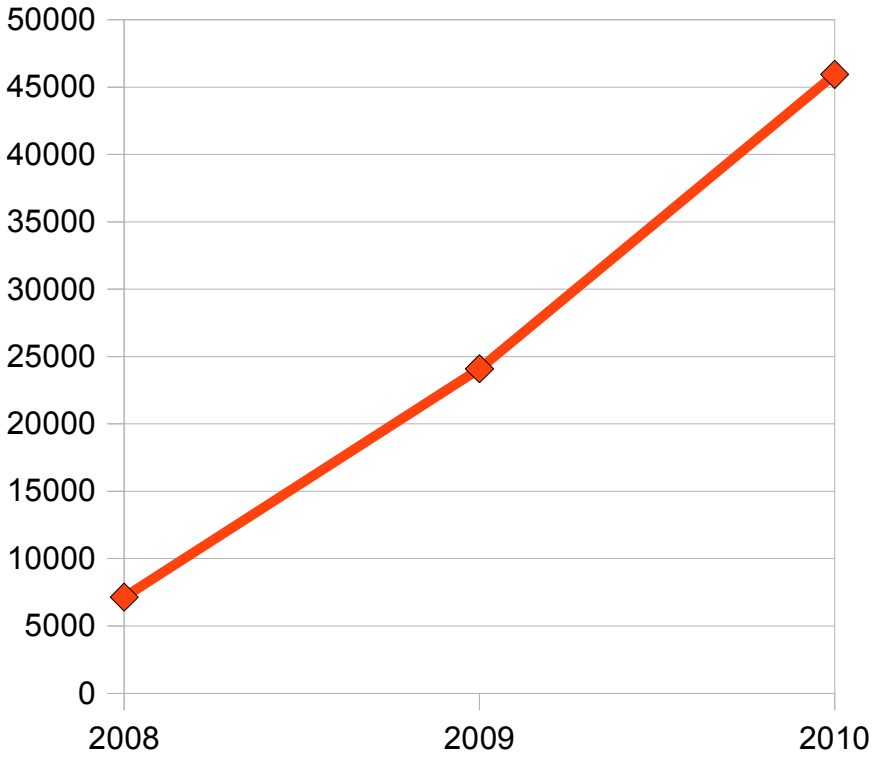
Luciferase reporter assays

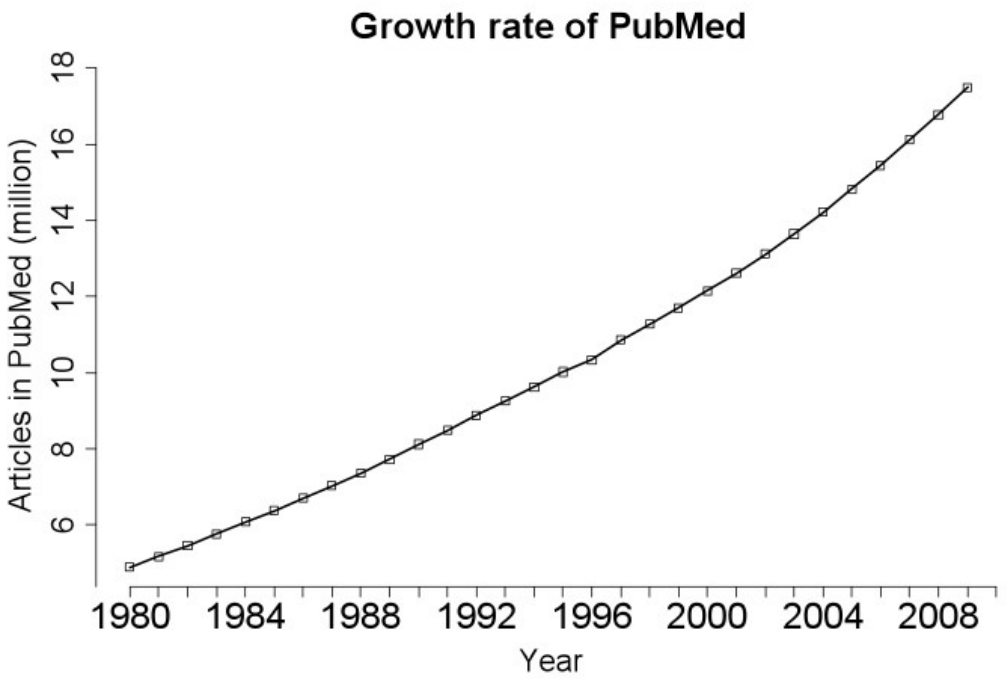
Neuro2a mouse neuroblastoma cells (0.8×10^{-5}) were transfected 24 h after plating with 200 ng of luciferase reporter plasmid DNA, 50 ng of pCDNA3.1-Arx-V5-His (131) or pCDNA3.1-V5-His (Invitrogen) and 50 ng of pRL-TK-*Renilla* luciferase plasmid DNA (Promega) using FuGENE 6 (Roche Diagnostics, Alameda, CA). Forty-eight hours post-transfection, cell lysis and measurement of *firefly* and *Renilla* luciferase activity was performed using the Dual-Glo Luciferase Assay System (Promega) according to the manufacturer's instructions using a Veritus Microplate Luminometer (Turner BioSystems, Sunnyvale, CA). Transfections were performed in quadruplicate, and three independent experiments were performed. The *firefly* luciferase activity was normalized according to the corresponding *Renilla* luciferase activity, and luciferase activity was reported as mean (±SEM) relative to pCDNA3.1-V5-His/luciferase transfection. Luciferase reporter constructs were generated using primer sequences identical to those used for ChIP, modified to include a *Hin*dIII on the 5′ end of the sense primer and a *Bam*HI site on the 3′ end of the antisense primer. These primers were used to PCR amplify product from E14.5 mouse genomic DNA, and the amplified products were cut with *Bam*HI and *Hin*dIII and cloned into the *Bam*HI and *Hin*dIII sites of PPREX3-TK-Luc (Addgene plasmid 1015) (133), which contains the thymidine kinase promoter upstream of luciferase, replacing the PPAR response element.

Five hundred nanograms of GST-Arx (286–430) was incubated with 0.1 pmol of labeled-probe in total 20 μl of binding buffer [20 mM HEPES pH 7.4, 50 mM KCl, 1 mM MgCl2, 1 mM DTT, 5% glycerol, 1 μg poly(dI-dC) (Roche)] for 30 min on ice. In competition experiments, 10 pmol cold probes were added. In antibody supershift experiment, 2 μl of anti-GST antibody (Amersham Biosciences) was added. The mixture was loaded on 5% polyacrylamide gel and electrophoresized (constant 33 mA) at 4°C in 1× TBE. Gels were dried and visualized on a phosphorimager (Molecular Dynamics). Oligonucleotides used were as follows: Ebf3: 5′-GCGATTTTCCCGATTAA TTAAAATATTAACGCA-3′ and 5′-GTGCGTTAATATTTT AATTAATCGGGAAAATCG-3′; Ebf3mut1: 5′-GCGATTTT CCCGATT**CC**TTAAAATATTAACGCA-3′ and 5′-GTGCG TTAATATTTTAA**GG**AATCGGGAAAATCG-3′; Ebf3mut2: 5′-GCGAT**CC**TCCCGATTAATTAAAATATTAACGCA-3′ and 5′-GTGCGTTAATATTTTAATTAATCGGGA**GG**ATCG -3′. Lmo1: 5′-GTAATGAATTGATTTAATTAACAGGGGA GTCTGA-3′ and 5′-GTCAGACTCCCCTGTTAATTAAATC AATTCATTA-3′; Lmo1mut1: 5′- GTAATGAATTGATTT**CC** TTAACAGGGGAGTCTGA -3′ and 5′-GTCAGACTCCCCT GTTAA**GG**AAATCAATTCATTA-3′; Lmo1mut2: 5′-GTAA TGAA**CC**GATTTAATTAACAGGGGAGTCTGA-3′ and 5′- GTCAGACTCCCCTGTTAATTAAATC**GG**TTCATTA-3′; Shox2: 5′-GCAAAATCCACGCTTAATTAAATTAATTAGG GA-3′ and 5′-GTCCCTAATTAATTTAATTAAGCGTGGAT TTTG-3′.

High-throughput versus in-depth analyses

Number of runs submitted to the NCBI short read archive





The identifiers

Luciferase reporter assays

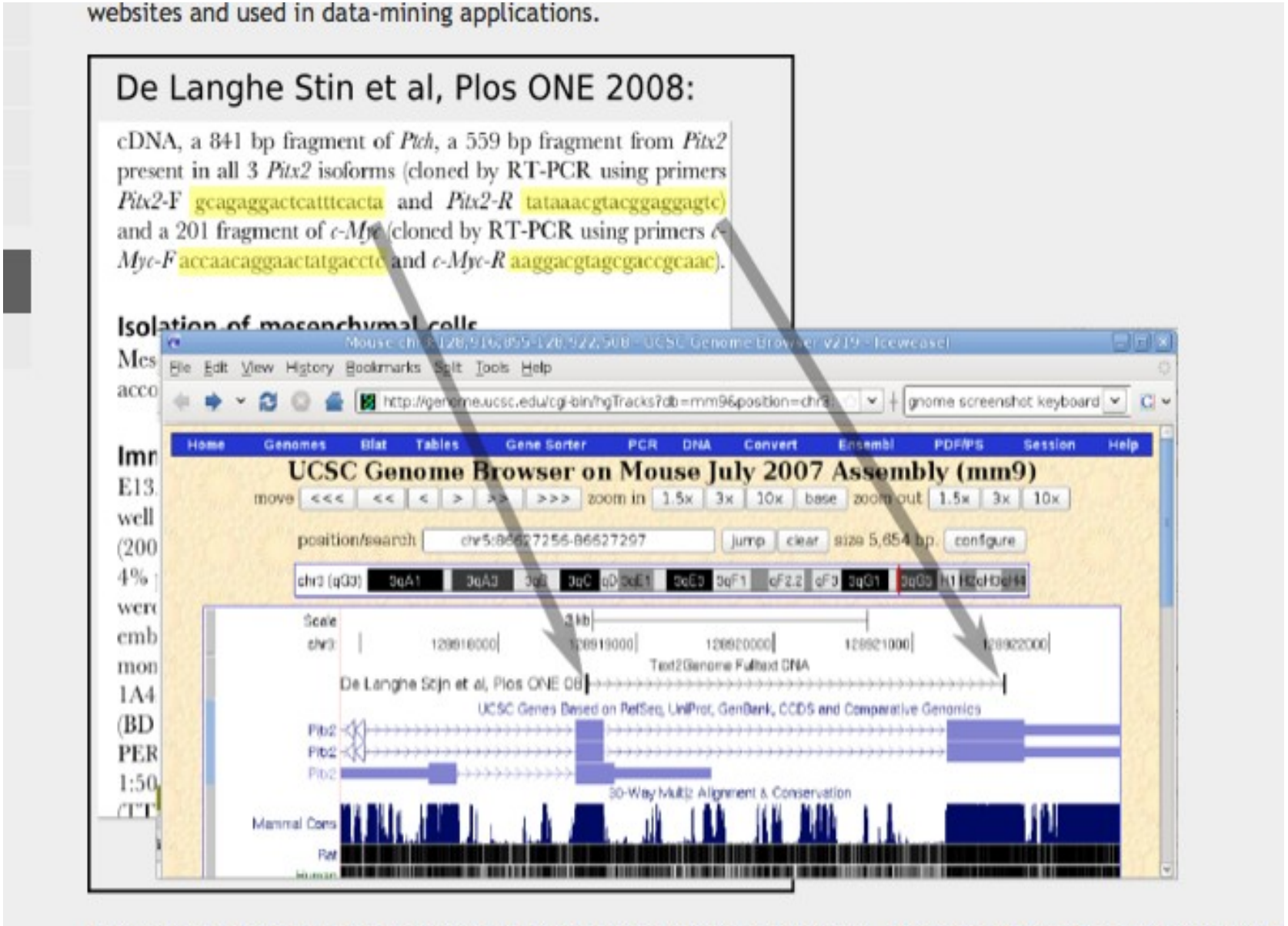
Neuro2a mouse neuroblastoma cells (0.8×10^{-5}) were transfected 24 h after plating with 200 ng of luciferase reporter plasmid DNA, 50 ng of pCDNA3.1-Arx-V5-His (131) or pCDNA3.1-V5-His (Invitrogen) and 50 ng of pRL-TK-*Renilla* luciferase plasmid DNA (Promega) using FuGENE 6 (Roche Diagnostics, Alameda, CA). Forty-eight hours post-transfection, cell lysis and measurement of *firefly* and *Renilla* luciferase activity was performed using the Dual-Glo Luciferase Assay System (Promega) according to the manufacturer's instructions using a Veritus Microplate Luminometer (Turner BioSystems, Sunnyvale, CA). Transfections were performed in quadruplicate, and three independent experiments were performed. The *firefly* luciferase activity was normalized according to the corresponding *Renilla* luciferase activity, and luciferase activity was reported as mean (±SEM) relative to pCDNA3.1-V5-His/luciferase transfection. Luciferase reporter constructs were generated using primer sequences identical to those used for ChIP, modified to include a *Hin*dIII on the 5′ end of the sense primer and a *Bam*HI site on the 3′ end of the antisense primer. These primers were used to PCR amplify product from E14.5 mouse genomic DNA, and the amplified products were cut with *Bam*HI and *Hin*dIII and cloned into the *Bam*HI and *Hin*dIII sites of PPREX3-TK-Luc (Addgene plasmid 1015) (133), which contains the thymidine kinase promoter upstream of luciferase, replacing the PPAR response element.

Five hundred nanograms of GST-Arx (286–430) was incubated with 0.1 pmol of labeled-probe in total 20 µl of binding buffer [20 mM HEPES pH 7.4, 50 mM KCl, 1 mM MgCl2, 1 mM DTT, 5% glycerol, 1 µg poly(dI-dC) (Roche)] for 30 min on ice. In competition experiments, 10 pmol cold probes were added. In antibody supershift experiment, 2 µl of anti-GST antibody (Amersham Biosciences) was added. The mixture was loaded on 5% polyacrylamide gel and electrophoresized (constant 33 mA) at 4°C in 1× TBE. Gels were dried and visualized on a phosphorimager (Molecular Dynamics). Oligonucleotides used were as follows: Ebf3: 5′-GCGATTTTCCCGATTAATTAAAATATTAACGCA-3′ and 5′-GTGCGTTAATATTTTAATTAATCGGGAAAATCG-3′; Ebf3mut1: 5′-GCGATTTTCCCGATT**CC**TTAAAATATTAACGCA-3′ and 5′-GTGCGTTAATATTTTAA**GG**AATCGGGAAAATCG-3′; Ebf3mut2: 5′-GCGAT**CC**TCCCGATTAATTAAAATATTAACGCA-3′ and 5′-GTGCGTTAATATTTTAATTAATCGGGA**GG**ATCG-3′. Lmo1: 5′-GTAATGAATTGATTTAATTAACAGGGGAGTCTGA-3′ and 5′-GTCAGACTCCCCTGTTAATTAAATCAATTCATTA-3′; Lmo1mut1: 5′-GTAATGAATTGATTT**CC**TTAACAGGGGAGTCTGA -3′ and 5′-GTCAGACTCCCCTGTTAA**GG**AAATCAATTCATTA-3′; Lmo1mut2: 5′-GTAATGAA**CC**GATTTAATTAACAGGGGAGTCTGA-3′ and 5′-GTCAGACTCCCCTGTTAATTAAATC**GG**TTCATTA-3′; Shox2: 5′-GCAAAATCCACGCTTAATTAAATTAATTAGGGA-3′ and 5′-GTCCCTAATTAATTTAATTAAGCGTGGATTTTG-3′.

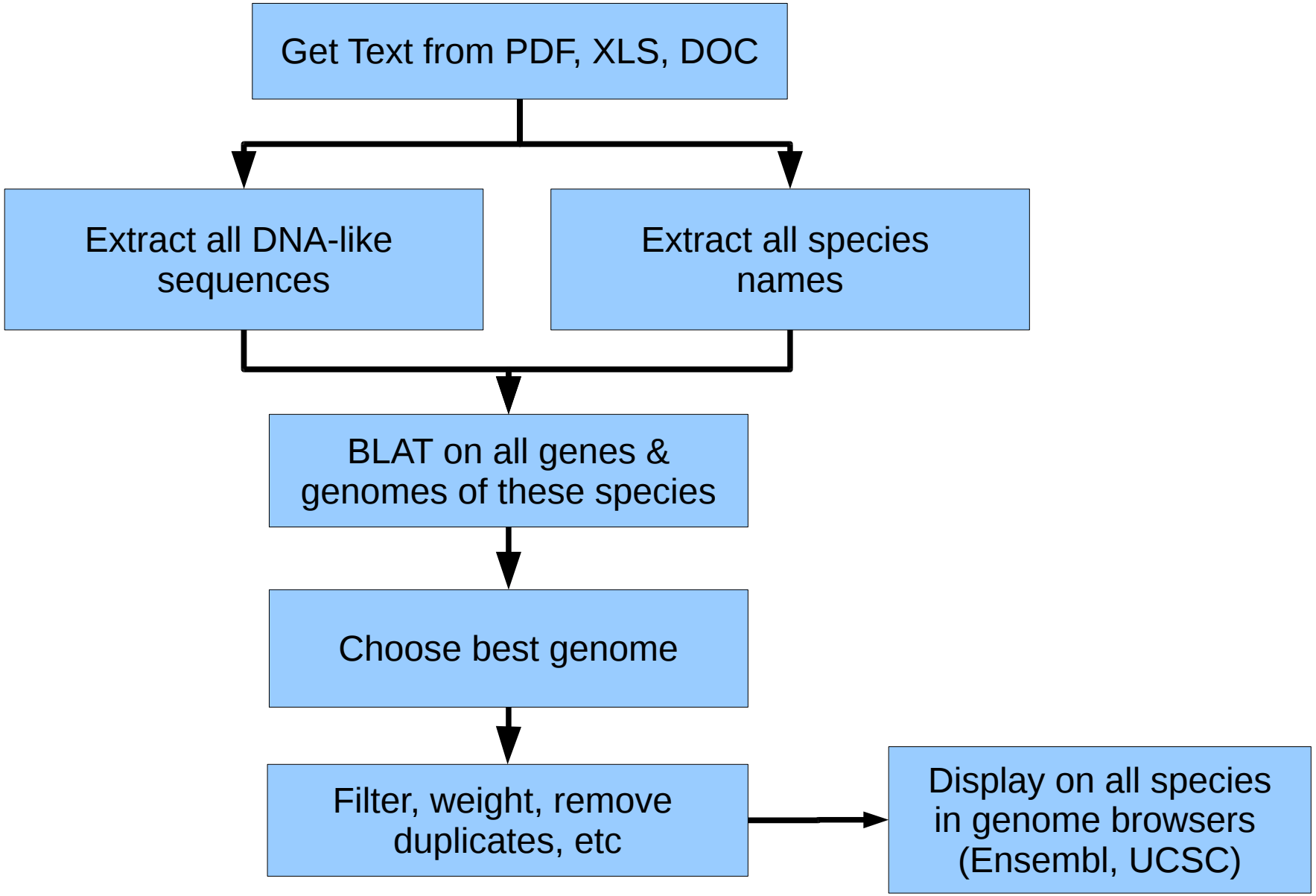
Eight sequences:

1- TCCAGTTCCCCAGTGTTTTACTAAGT
2 - GCTCTTGGCCATTAATCCAGGATT
3- TAAGCTAATGGCGGGCACCT
3 - CTCGCTCTCACCAGAGTGCA
5- CCGTAATGGATTTTGAGATGGGA
6 - TGAATTGGTGGTGTGTGTGC
7- TACTCGCGGCTTTACGGG
8 - TGGAACAGGGAGGAGCAGAGAGCA

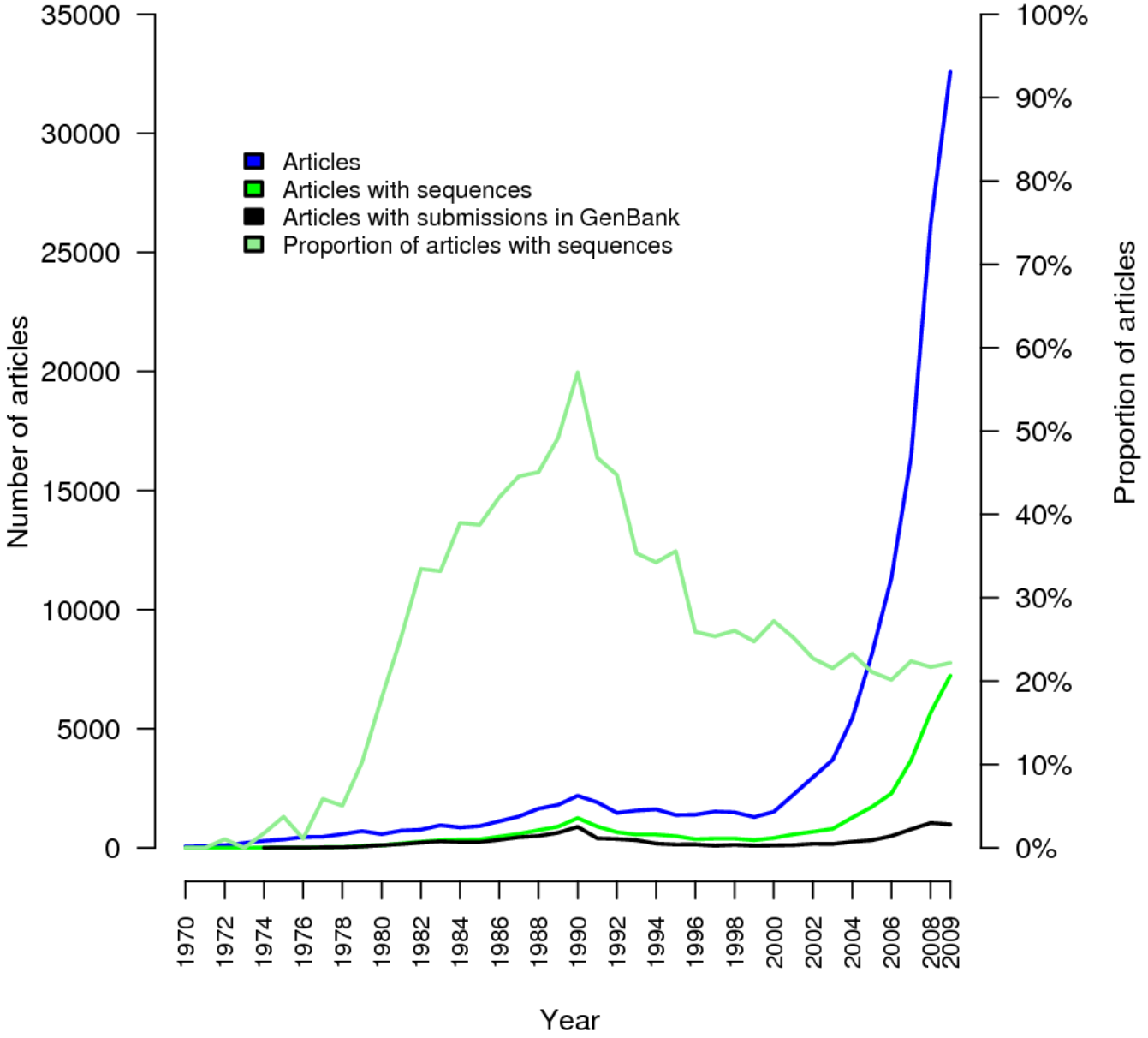
Text2Genome.org



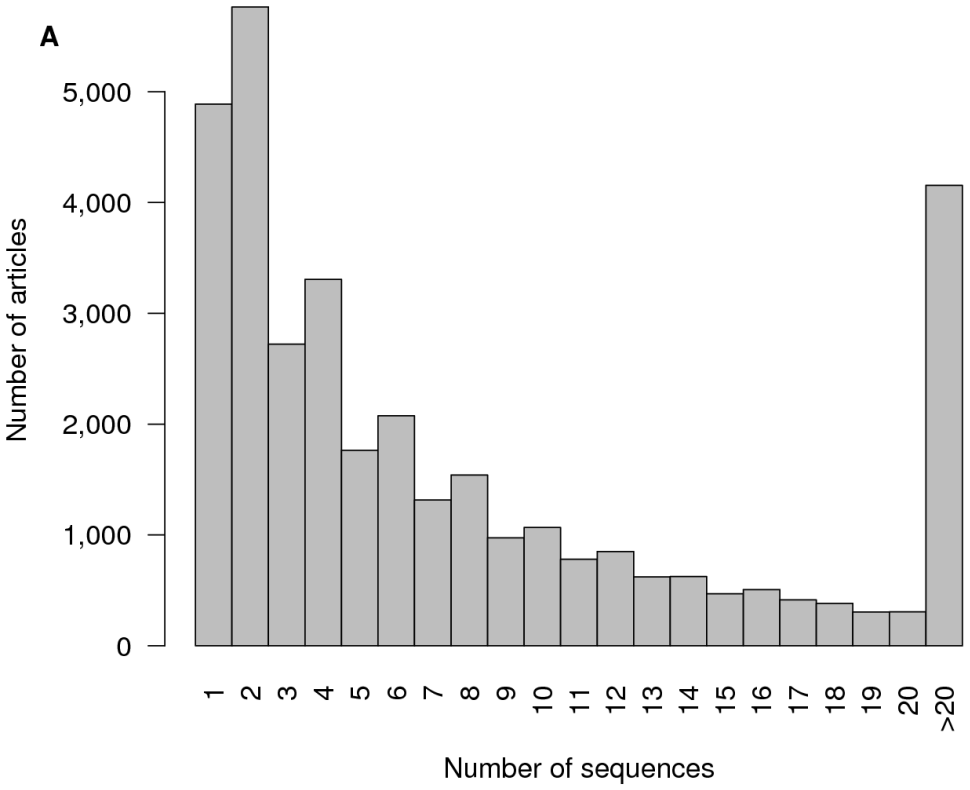
Text2Genome.org Pipeline

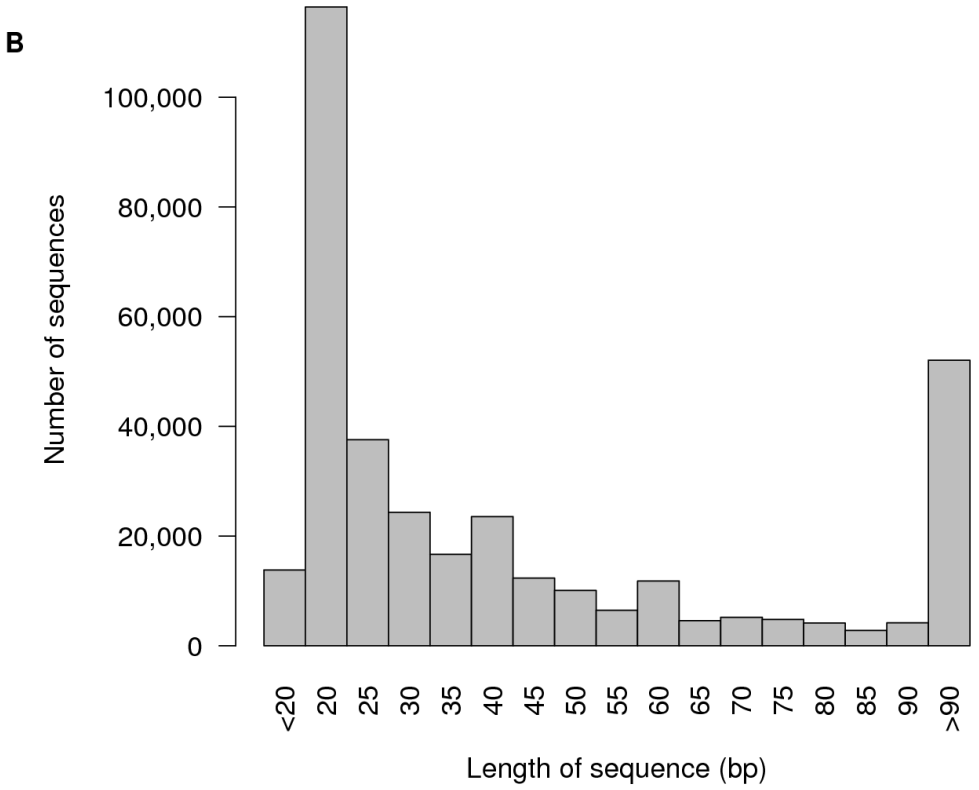


Applied to PubMed Central Open Access Subset (150k articles)

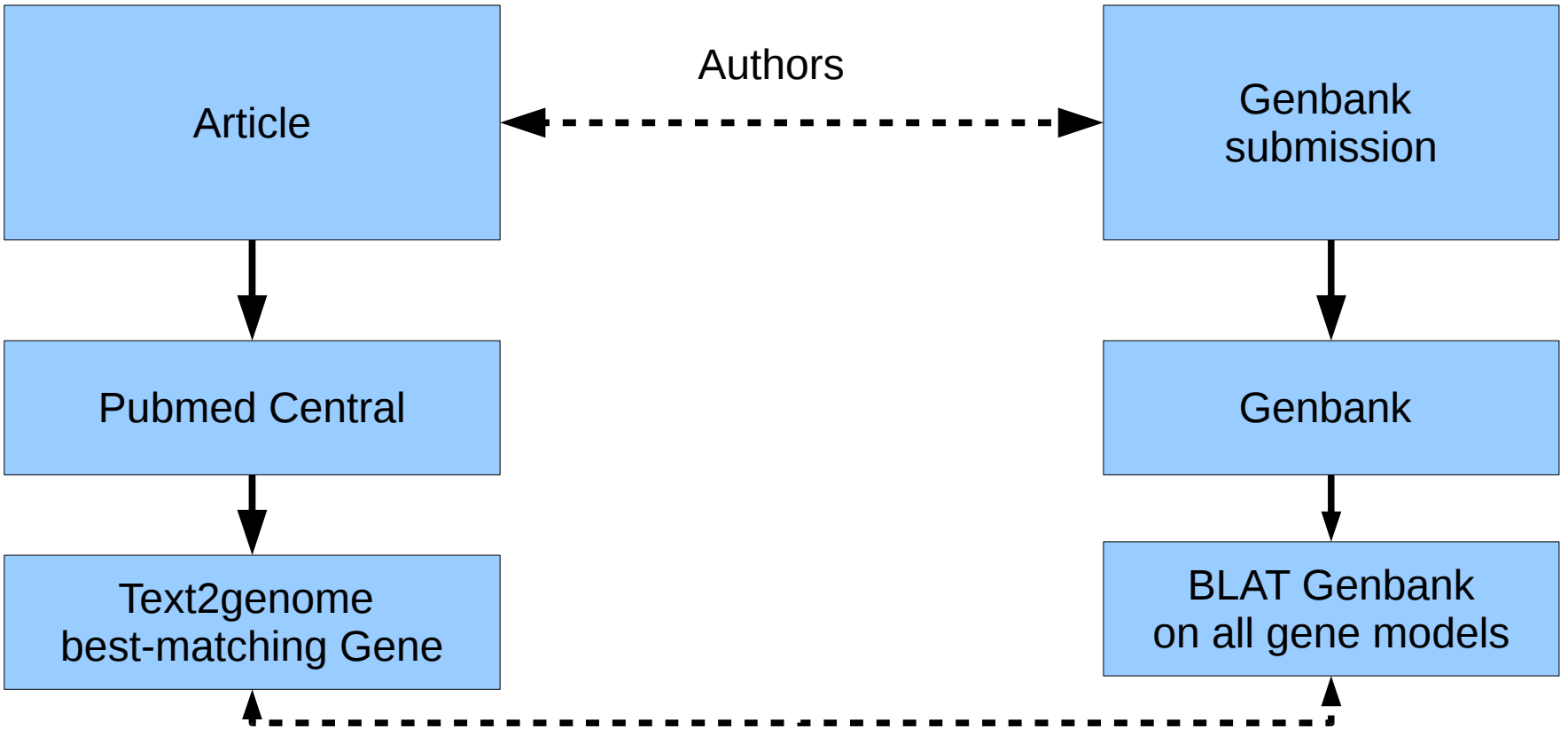


Most sequences are primers





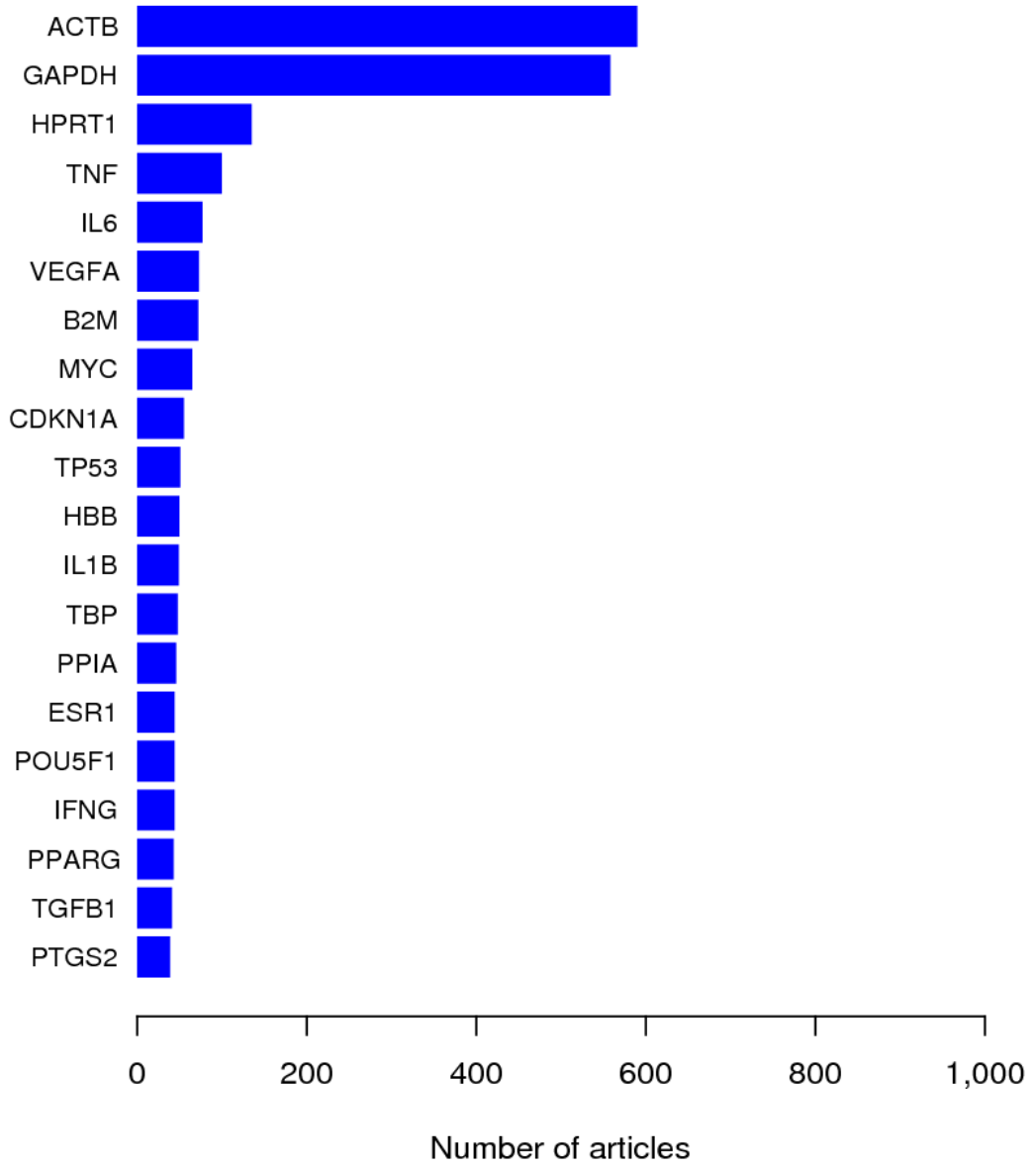
Estimate Precision vs. Genbank



Predict correct species: 95%

Predict correct gene: 88%

Most frequent genes with sequences in articles are RT-PCR loci



Technical details

- Parsing from PDF/XML/TXT
- And supplemental files (XLS/DOC/...)!
- DNA sequences look like...?
- How do you quickly search for 150k species names?
- Short sequences
- 100GB of genome data
- Display: Ensembl DAS Server, UCSC custom tracks

www.text2genome.org

Human Molecular Genetics, 2008, Vol. 17, No. 23 *3740–3760*
doi:10.1093/hmg/ddn271
Advance Access published on September 16, 2008

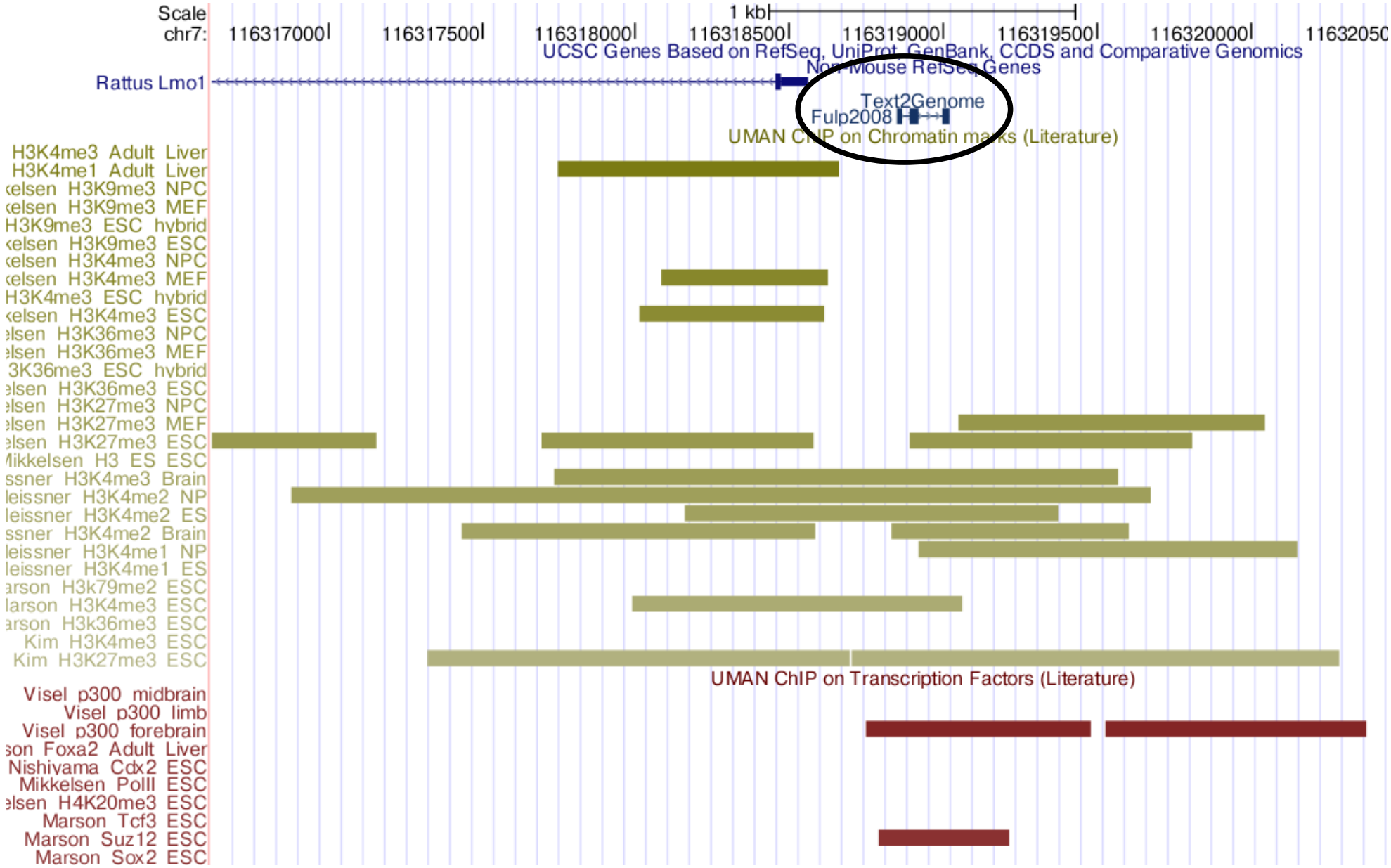
Genome Annotation: Links to best and chained genome matches

SeqNo	Coordinate Range
7, 8, 28, 29	chr3:66786225-66786328
9, 10, 22, 23, 26, 27	chr7:116318846-116319016
31	chr7:69524512-69524547
13	chr6:125115600-125115624
4, 5	chrY:1917607-1920436
0, 1	chrX:90533156-90535654
13	chr12:5597479-5597862
13	chr18:55998656-55998680
11, 12, 16, 17, 18, 19, 20, 21	chr7:144507440-144507660

Recognized sequences in fulltext

SeqNo	file name	Recognized DNA
0	PMC2581427.pdf	TGGAGCGGGGACAGGGGTGAGGTT
1	PMC2581427.pdf	GGCCGGTCTCTTTCTTTCTACTCA
2	PMC2581427.pdf	CAATGCTGTTTCACTGGTTATG
3	PMC2581427.pdf	CATTGCCCCTGTTTCACTATC
4	PMC2581427.pdf	CAGAAATGAACTACTGCATCCC
5	PMC2581427.pdf	AACTTGTGCCTCTCACCACG
6	PMC2581427.pdf	AAGAACCCCAAAAGCTAAG
7	PMC2581427.pdf	TCCAGTTCCCCAGTGTTTTACTAAGT
8	PMC2581427.pdf	GCTCTTGGCCATTAATCCAGGATT

Is there any literature analyzing this genomic region?



Future directions

- Standard UCSC genome browser track
- More articles to come:
  - Elsevier (30 % of Pubmed-abstracts)
  - Nature Publishing
  - **...your journal... ?**

MANCHESTER 1824
The University of Manchester

Acknowledgements

- Casey Bergman (Manchester), Jean-Stephane Joly (Gif-sur-Yvette, France):
  - Motivation & support for an idea which sounded crazy some years ago and took some time to realize
- Martin Gerner (Manchester):
  - Fast species name recognizer: http://linnaeus.sourceforge.net
- Hiram Clawson (UCSC):
  - Integration into UCSC browser (coming soon)
- Funding: EU 7th Framework Project CISSTEM (www.cisstem.eu), BBSRC Project “pubmed2ensembl” (www.pubmed2ensembl.org)