Visualizing Whole Genomes

The UCSC Human Genome Browser: Hands-on Exercise

What do you do with a whole genome sequence once it is complete? Most genome-wide analyses require having the data, but not necessarily visualizing the assembled genome. Looking at a well-assembled, well-annotated genome, however, can be very interesting and can help biologists both test and develop new hypotheses. Genome browsers offer a visual portal for whole genome data exploration and analysis. They are useful for looking at individual genes in their genomic neighborhood, comparative genomics, and for integrating many types of bioinformatic data in one place. Most whole genome sequencing projects intend, at some point, to make the genome they have sequenced publicly-available in browser form. Very few species' genomes have been well assembled and annotated, thus there are only a few good browsers currently available (e.g., NCBI's Map Viewer, Ensembl, and the UCSC Genome Browser; linked from the web page).

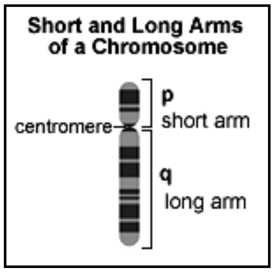
Overview: Using genome browsers can be helpful, even if you are studying a specific gene or gene family, you may want to study it in a genome-wide context. Today, we will explore the genes coding for amylase, an important enzyme that breaks down starch into sugar. Amylase is the major protein found in saliva, and it is also excreted in the pancreas. Salivary amylase (coded for by AMY1 genes) is one of the most extensively studied proteins in humans, whereas relatively little is known about pancreatic amylase (coded for by AMY2 genes). Today, we will use a combination of web tools and genome browser tools to learn about this gene family, and then develop a testable hypothesis to learn more about functional genetics and genomics.

Let's start our investigation by navigating to the **GQuery** page (previously known as **Entrez**), a site where we can search across all NCBI databases with a single search term (http://www.ncbi.nlm.nih.gov/sites/gquery). We want information on amylase, however this gene family is ancient. That is, an amylase gene was present in the ancestor of both prokaryotes (bacteria) and eukaryotes, thus there are homologues of amylase in a wide range of taxa and we need to specify that we want to find out about *human amylase*. Type "**human and amylase**" in the query window, select **Search** and look at the hits you get in ALL the different databases. Amylase was the first enzyme discovered and isolated, and has been the subject of intense research because it is hypothesized that shifts toward starch-rich diets have likely played a large role in human evolution.
*How many papers on human amylase are indexed in **Pubmed**?* ________

Halfway down the page, you will see hits in the **Gene** database, click on the **Gene: collected information about gene loci** link and browse the *entire* first page of results. Look at the gene titles and the information listed below the title. Identify 5 genes that are *clearly* human amylase genes based on their names. There is also a 6th gene that may be relevant to our exploration today—an amylase pseudogene. Find this gene as well, and add it to your list.

List your 6 genes here (use the symbol, e.g., AMY1A, don't write out the whole name).

______________________ ______________________
______________________ ______________________
______________________ ______________________



If you click on the gene name for the first hit, **AMY1A**, you will be redirected to the page for that gene specifically. Here, you can read basic information about the gene, including its function and location. *HINT*: Look under headers **Genomic context** and **Summary**.

What chromosome and what arm (p or q) are all the amylase genes located on? ________

Where is this amylase gene mainly expressed? ______________________________

Record the bp location (begin-end) for AMY1A: ______________________________

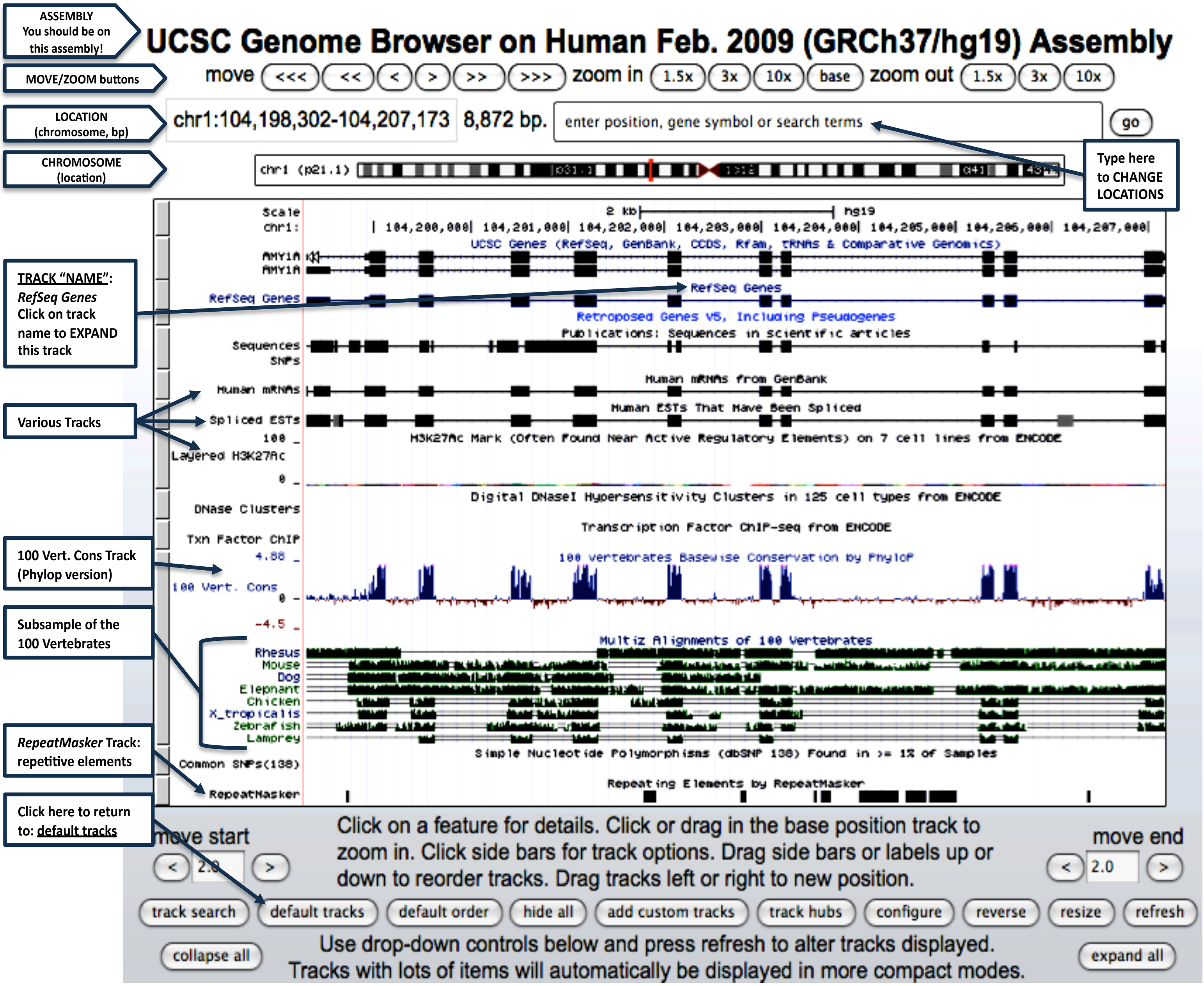
On the right hand side of the page, there are links to table of contents, tools, information and other resources. Under the **Link to other resources** header on the right, click on the **UCSC** link, which takes you directly to the amylase gene in the **UCSC** genome browser. A new window will open containing a title named either: "**refGene**" or **"RefSeq Genes"** with a few links below it. Click the top link.

You are now in a genome browser! The first (title) line of UCSC browser has information on what version of the human genome assembly that you are looking at. Typically, by default, the assembly should be the latest and most complete, however you may not have been directed to the most recent assembly. The top line should read: **UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly.** If you are ***NOT*** on this assembly, complete the instructions on the following page:

Click on **Genomes** tab in the menu at the top of the page. This will direct you to a new page, under the header **Human (*Homo sapiens*) Genome Browser Gateway**, select the link: "**Click here to reset the browser user interface settings to their defaults**", to refresh your browser to have the most recent assembly. To return to the genome browser page click on the **Genome Browser** tab in menu at the top of the page (the assembly should now be the ***Feb. 2009*** version). *NOTE: it is very important to know which genome assembly you are looking at.* At this point you will need to reorient yourself within the genome browser back to chromosome 1 (where the amylase gene is located). To do this type "**amylase**" in the search box at the top of the page, and select **Go**. You will be redirected to a new page with a list of gene links; under **RefSeq Genes** click on this link: **AMY1B at chr:104198325-104207172** (should be near the bottom of the list). This will redirect you to the genome browser page.

You can now continue to navigate through the genome browser page. **Refer to the annotated figure below to orient yourself within the browser page.** The second line has buttons for navigating within the browser (**move** left and right; **zoom in** and **zoom out**). The third line shows you where you are, and how much of the genome you are currently looking at (this is important!). In the fourth line, there is a cartoon of the chromosome you are in, with a red box showing you exactly where you are (remember where the location of the amylase gene was according to the **Gene** record? It should be the same.)

What chromosome are you on? ___________ *How many bps are visible in your window?* ___________



UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

chr1:104,198,302-104,207,173 8,872 bp.

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

Use drop-down controls below and press refresh to alter tracks displayed. Tracks with lots of items will automatically be displayed in more compact modes.

ASSEMBLY You should be on this assembly!

MOVE/ZOOM buttons

LOCATION (chromosome, bp)

CHROMOSOME (location)

Type here to CHANGE LOCATIONS

TRACK "NAME": *RefSeq Genes* Click on track name to EXPAND this track

Various Tracks

100 Vert. Cons Track (Phylop version)

Subsample of the 100 Vertebrates

RepeatMasker Track: repetitive elements

Click here to return to: default tracks

move <<< << < > >> >>>
zoom in 1.5x 3x 10x base
zoom out 1.5x 3x 10x
enter position, gene symbol or search terms
go
chr1 (p21.1)
p31.1
1q12
q41
q43
Scale
2 kb
hg19
chr1:
104,200,000
104,201,000
104,202,000
104,203,000
104,204,000
104,205,000
104,206,000
104,207,000
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
AMY1A
AMY1A
RefSeq Genes
RefSeq Genes
Retroposed Genes V5, Including Pseudogenes
Publications: Sequences in scientific articles
Sequences
SNPs
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
100
Layered H3K27Ac
0
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Digital DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE
DNase Clusters
Transcription Factor ChIP-seq from ENCODE
Txn Factor ChIP
4.88
100 vertebrates Basewise Conservation by PhyloP
100 Vert. Cons
0
-4.5
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
Simple Nucleotide Polymorphisms (dbSNP 138) Found in >= 1% of Samples
Common SNPs(138)
Repeating Elements by RepeatMasker
RepeatMasker
move start
< 2.0 >
move end
< 2.0 >
track search
default tracks
default order
hide all
add custom tracks
track hubs
configure
reverse
resize
refresh
collapse all
expand all

The large window below the chromosome cartoon contains multiple "tracks". Each track is described by a short name *above* the track, with more details on the left-hand side. Clicking on the track name will ***expand*** the track, and clicking on data within a track will ***take you to another page*** with the details about the data. In the main browser window (with **default settings**) there are a number of tracks showing. Scroll up and down to get acquainted with the data available. If you scroll to the bottom of the page, you will notice that below the main window, there are many, many more options for tracks that you can add to your window. In fact, you can upload your own data to create custom tracks so you can analyze data you generate against publicly available data.

Currently, you should be looking at about a ~9000 bp-long region of the chromosome. Look at the top few tracks and their labels on the left hand side. Before going further, notice:

(1) The top few tracks show you cartoons of the genes; they have boxes connected by thin lines (as with most cartoons of genes, boxes represent the exons and thin lines represent the introns).

(2) The **AMY1A**, **AMY1B**, and **AMY1C** genes appear as individual tracks at the *same* location.
Confused? You should be! In the **Gene** database, they were represented as different loci, but here they appear to be different genes that are located in exactly the same place. Hopefully, we will clear up this confusion by learning more about this gene family.

Look at the **Human mRNAs from GenBank** track. Notice that the sequence data used to build this track shows that mRNAs that were sequenced match up almost perfectly with the exons described in the genome sequence (DNA). Below this, notice a similar track: **Human ESTs That Have Been Spliced**; ESTs (expressed sequence tags) are likely candidates for mRNAs, but they haven't been validated as mRNAs yet.

Look at the track names on the left-hand side bar, and jump down a few tracks to a track called **100 Vert. Cons.** This track has two interchangeable names; make sure you are on the **100 Vertebrates Basewise Conservation by Phylop** view. If you are on the **Vertebrate Multiz Alignment & Conservation (100 Species)** view, just click on the track name to switch to the **Phylop** version. In the **Phylop** version of the **100 Vert. Cons** track: the data are different—the higher the peak, the greater the level of conservation in an alignment of amylase genes from 100 different species of vertebrates, a subsample of which are shown below this track. As you would expect for a gene such as amylase with a widespread and functional locus, the exon regions (which are coding sequence) are highly conserved across species. However, the level of conservation (or sequence constraint) on intronic regions is much lower.

Look at the individual species tracks below, located under the track name **Multiz Alignments of 100 Vertebrates**. The dark bars indicate identity at a given nucleotide position. Notice that **rhesus monkey** is nearly identical at every position to human (you may also notice some gaps in the rhesus alignment, these are likely the result of a rearrangements or indels). Even lamprey (the jawless fish, with which we have not shared an ancestor for over 500 million years), shown on the bottom of this set of tracks has high sequence similarity in exonic regions at this locus.

Zoom OUT 10x, so approximately 88,000 bp are now represented in your window. If you have added or collapsed tracks, you can scroll to the bottom of the browser window and press the **default tracks** button to return the window to default settings: **Do this now**! Resetting to default tracks will collapse the **RefSeq Genes** track, for the following steps ***re-expand*** this track (click on the track name).

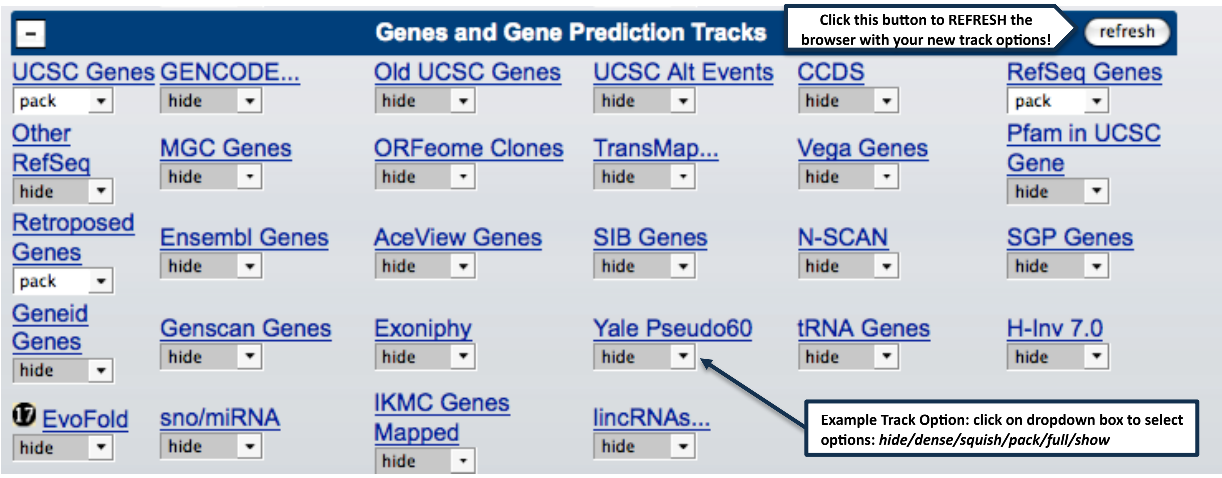
Look at the bottom track (**RepeatMasker**) where the dark lines do not represent sequence identity, but instead represent repetitive sequence. *What do you notice about the relationship between the data represented in the top tracks (which depict genes) and this bottom track?*

____

Zoom OUT 3x, so approximately 266,000 bp are now represented in your window. Locate the **RefSeq Genes** track near the top of the list (make sure this track is expanded).
*How many **different** spots along the chromosome have salivary amylase genes (AMY1) been annotated?* ________
How about pancreatic amylase genes (AMY2)? ________

You are probably starting to get the feeling that this gene is prone to gene duplication. In fact, it looks as if there was a recent duplication of AMY1 (generating the 3 AMY1 loci), and that there may have been a more ancient duplication that led to the formation of the AMY1 and AMY2 genes (AMY2 also now appears to have duplicated at least once) based on the browser data.

OPTIONAL TRACKS (located below the genome browser graphics window):



Click this button to REFRESH the browser with your new track options!

Example Track Option: click on dropdown box to select options: *hide/dense/squish/pack/full/show*

UCSC Genes pack	GENCODE... hide	Old UCSC Genes hide	UCSC Alt Events hide	CCDS hide	RefSeq Genes pack
Other RefSeq hide	MGC Genes hide	ORFeome Clones hide	TransMap... hide	Vega Genes hide	Pfam in UCSC Gene hide
Retroposed Genes pack	Ensembl Genes hide	AceView Genes hide	SIB Genes hide	N-SCAN hide	SGP Genes hide
Geneid Genes hide	Genscan Genes hide	Exoniphy hide	Yale Pseudo60 hide	tRNA Genes hide	H-Inv 7.0 hide
17 EvoFold hide	sno/miRNA hide	IKMC Genes Mapped hide	lincRNAs... hide		

Scroll below the genome browser window, to the lists of optional tracks and find the **Genes and Gene Prediction Tracks** section. Locate the track called **Yale Pseudo60** and select **full**, and then hit the **refresh** button at the top right of the section. Now go back to the browser window to view this track (titled **Yale Pseudogenes based on Ensembl Release 60**) and make sure approximately 240,000 bp are now represented in your window (zoom in/out if necessary). Look for the graphic representing the largest pseudogene (**PGOHUM00000244824**) —this is the one we saw annotated in the **Gene** database at the beginning of the module. ***HINT***: if you scroll your cursor over the pseudogene boxes, the name will appear.

Is there evidence from the tracks depicting gene expression that this pseudogene is expressed? ________
Is the level of conservation (look at the 100 Vert. Cons track) at the pseudogene locus higher or lower than the "true" amylase genes? _______ *Why is this?* ___________________________________

You can imagine that when biologists were first studying this gene (before the whole genome was sequenced), it may have been pretty confusing to have 5 copies of a very similar gene (plus, in a diploid organism, there are another 5 versions on the other chromosome 1 homologue). In fact, it turns out the amylase gene family is even more variable than what is represented here! Luckily, there is a track to show us data from studies looking at copy number variation among different individuals.

Zoom OUT 10x again, so you are now looking at 2,600,000 bps. Return the window to default tracks if you need to. You should be able to see a gene to the *left* of the amylase cluster now because you are looking at a large section of DNA. *What gene is neighboring the amylase cluster?* ____
What is its function? (HINT: click on one of the gene exons) ____

Scroll down to the track option window labeled **Variation**, and select the **DGV Struct Var** (pull down and select **squish** so you can see the track, but not every piece of data). Hit **refresh** on the right, and then scroll up and look for the new tracks you've added in your browser.

Under the **Database of Genomic Variants** track (displays two similar sub-tracks) you can now see tons of data showing cases where the sequences of different individuals show copy number variation at a given locus in the genome. Red bars indicate cases of copy loss, blue bars show cases of copy gain, and brown bars indicate copies have both been gained or lost in that sample or study. Look at the region below both the *amylase cluster* and the *collagen gene* in your current view, and gauge the amount of copy number variation at these two loci.

Which gene, collagen or amylase, appears to have greater copy number variation among individuals?
____

Does this clear up some of the confusion over why there might be multiple *similar* or *identical* genes annotated for a given gene that do not clearly correspond to a specific region? So, instead of remaining confused about the dynamic copy number variation of the amylase gene, lets use this variation problem as an opportunity. Now that you have identified amylase as a highly variable copy number gene family, develop an ***hypothesis*** that you could test to determine what factors might underlie copy number variation in this gene family.

First, think about what more copies of the amylase gene might mean for digesting starchy foods:

How would you test this? What samples and what methods would you use? (Discuss with your partner).

Second, think about populations that you are familiar with (perhaps different groups or regions in East Africa with different typical levels of starch in their diet). Formulate a hypothesis about which groups you would expect to have higher copy number of amylase genes:

How would you test this? What samples and what methods would you use? What would your control genes or populations be? (Discuss with your partner).

Explore more tracks on your own…..

What tissue type is AMY1A most expressed in?

HINT: Use the **GNF Atlas 2** track to see expression data for several tissues based on microarray data; does the data corroborate what you know about where amylase is expressed? Click on one of the colorful boxes for a tissue of interest to obtain more detailed information about the expression levels in this tissue, and to view the **color key** (very useful for interpreting the browser display!) Remember salivary glands and the pancreas are both categorized as ***glands*** not organs.

Alu transposons are plentiful repeat sequences found in the human genome (over 1 million copies!).
Are there any Alu elements in this locus?

HINT: Under the **Repeats** track options, change **RepeatMasker** to **full** view and **refresh.** Expand the track to explore in detail, zoom in/out so that ~9,000 bp are shown in the window. Alus are in the category of repeats known as **SINEs** and if you scroll over individual SINEs in that track you will see the name of the element.

This UCSC website has many more resources than the genome browser (they are linked at the top of the page). For example, look under the **Tools** menu. From here, it is possible to do **"*in silico*" PCR** or look at *in situ* data (**Visigene**), and much much more…feel free to explore!

Further Reading:

*1. C. Alkan, B. P. et al., Applications of next-generation sequencing: Genome structural variation discovery and genotyping. Nat Rev Gen 12, 363-375 (2011). 2. A. S. Daar, et al., Ethics watch - Implications of copy-number variation in the human genome: A time for questions. Nat Rev Gen 7, 414-414 (2006). 3. A. L. Mandel, et al., Individual Differences in AMY1 Gene Copy Number, Salivary alpha-Amylase Levels, and the Perception of Oral Starch. Plos One **5**, (2010). 4. J. Novembre, A. Di Rienzo, Spatial patterns of variation due to natural selection in humans. Nat Rev Gen **10**, 745-755 (2009). 5. M. O'Bleness, et al., Evolution of genetic and genomic features unique to the human lineage. Nat Rev Gen **13**, 853-866 (2012). 6. F. G. Oppenheim, et al., in Oral-Based Diagnostics, D. Malamud, R. S. Niedbala, Eds. (2007), vol. 1098, pp. 22-50. 7. G. H. Perry, et al., Diet and the evolution of human amylase gene copy number variation. Nat Gen **39**, 1256-1260 (2007). 8. S. Ruhl, The scientific exploration of saliva in the post-proteomic era: from database back to basic function. Expert Review of Proteomics **9**, 85-96 (2012). 9. M. Skipper, Genomics - Copy number variation map. Nat Rev Gen **8**, 2-2 (2007). 10. B. T. Squires, Human salivary amylase secretion in relation to diet. Journal of Physiology-London **119**, 153-156 (1953).*