

Sequence Capture Experiment Design

Missy Hawkins
PhD Candidate



Sequence Capture

- Essentially reducing genomic datasets from entire genomes to smaller subsets, for many individuals
- Can be used for any scale genomic question
- Decreases cost for many individuals (as compared with shotgun sequencing)
- Better for use on degraded samples than some HTS methods (RAD-seq)



Project Goals

- Large Scale Phylogenies:
 - ‘Universal markers’ eg. UCE’s
 - Exons/Introns
 - Chromosomes or genes etc.
- Small Scale/ population studies:
 - SNP’s
 - Species ID’s
 - Etc.



A priori Information

- Genomic resources available
 - Previous genetic studies done?
 - All published information can be used
 - Markers found un/informative?
 - Metrosideros plants: ended up generating similar results to microsatellites from ~5,000 SNPs
 - Introns and Exons with previous knowledge useful



Practical Concerns

- How many samples total for project?
- What sequencing platform and indices?
- Will this tool be used for additional projects?
- Which kit best suits your needs?

– Example calculation:

(based on MYbaits 20,000 probe kit for 12 samples-smallest kit available)

Kit cost: \$2400 + S&H = \$2,450

If 5,000 SNPs included then during synthesis of a 20k kit repeated many times (concentration of 500ng/μl)

Dilute probe set 1:10 stretches from 12 'samples' (can also multiplex) to 12*10=120 samples (\$20/sample)

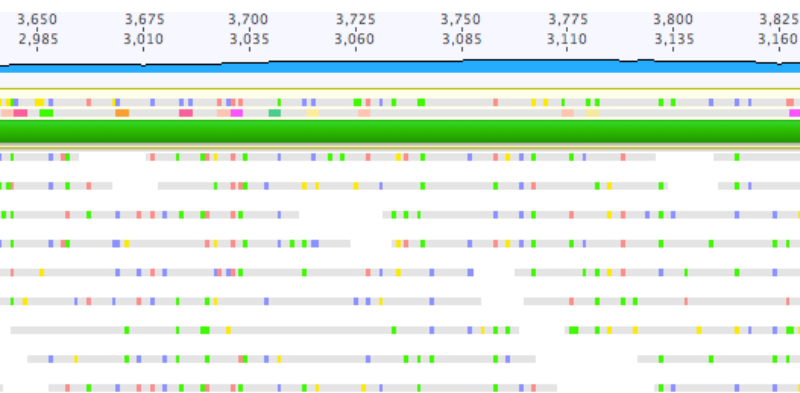
If multiplexed with just 2 individuals = 120 * 2 = 240 individuals per kit or \$10 per sample (2400/240) versus \$200/sample as published on website

SeqCap EZ Human Exome Library v3.0



In-Solution Design Steps

- Tiling
 - 2x usually plenty
- Number of probes
 - 4,000 or 40,000?
- Dilution of probes
 - Can dilute up to 1:15 but unstable after further dilution
 - must dilute at each enrichment



Pricing for Custom Kits				
Kit Name (Maximum # of bait sequences)	MYbaits-1 (20,000)	MYbaits-2 (40,000)	MYbaits-3 (60,000)	MYbaits-10 (200,000)
# of Captures	Price per Kit (USD)			
12	\$2,400	\$3,000	\$3,600	\$7,200
24	\$3,600	\$4,500	\$5,400	\$10,800
48	\$5,760	\$7,200	\$8,640	\$17,280
96	\$8,640	\$10,800	\$12,960	\$25,920
192	\$13,440	\$16,800	\$20,160	\$40,320
384	\$23,040	\$28,800	\$34,560	\$69,120
768	\$38,400	\$48,000	\$57,600	\$115,200
# of Captures	Effective Price Per Capture (USD)			
12	\$200	\$250	\$300	\$600
24	\$150	\$187.5	\$225	\$450
48	\$120	\$150	\$180	\$360
96	\$90	\$112.5	\$135	\$270
192	\$70	\$87.5	\$105	\$210
384	\$60	\$75	\$90	\$180
768	\$50	\$62.5	\$75	\$150

Target size per 20,000 baits depends on the bait length, tiling density and number of target loci. Please consult with us for more details.

Practical: EctoBaits

Simultaneous identification of host, vector and pathogen DNA via in-solution capture











Melissa T. R. Hawkins*, **Michael G. Campana***, Kristin Stewardson, Justin Lock, Kristofer M. Helgen, Hillary Young, Leah Card Jesús E. Maldonado, William J. McShea, Robert C. Fleischer

In Prep for Molecular Ecology Resources

- Project Goal: ID of host, vector and pathogen from tick samples
- Workflow:
 - Used mostly published Genbank sequences
 - Align sequences
 - Reduce redundancy (if necessary)
 - Split reads to probe lengths
 - Spend some time explaining project to Tech Support and send files for QC
 - Will test for base composition, strings of nucleotides and ambiguities etc.
 - Wait 6-8 weeks and test array!

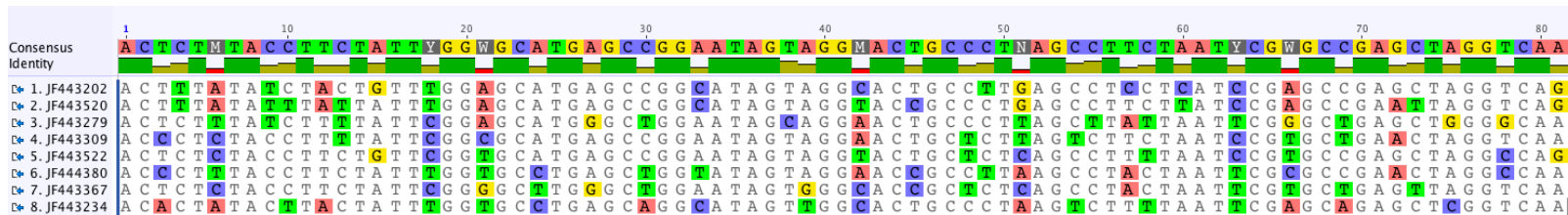
Download Genbank Sequences

- This project included African and American Mammals, Birds, Ticks, Pathogens
 - This study used primarily CO1 and Cyt *b*

	• AB004237	Felis catus mitochondrial DNA for cytochrome...	Felis catus	1,140	AB004237	2575782	DNA
	• AB015077	Sus scrofa domesticus mitochondrial cytb gen...	Sus scrofa	1,140	AB015077	3241885	DNA
	• AB015081	Sus scrofa domesticus mitochondrial cytb gen...	Sus scrofa	1,140	AB015081	3241893	DNA
	• AB462161	Procyon lotor mitochondrial CYTB gene for cyt...	Procyon lotor	1,140	AB462161	347582092	DNA
	• AF007908	Ursus americanus cinnamomum cytochrome b...	Ursus amer...	719	AF007908	2305025	DNA
	• AF007934	Ursus americanus americanus cytochrome b (c...	Ursus amer...	719	AF007934	2305077	DNA
	• AF028140	Canis latrans cytochrome b (cytb) gene, mitoch...	Canis latrans	396	AF028140	2826650	DNA
	• AF028156	Urocyon cinereoargenteus cytochrome b (cytb)...	Urocyon cin...	396	AF028156	2826682	DNA
	• AF057121	Lontra canadensis cytochrome b (cytb) gene, m...	Lontra cana...	1,140	AF057121	3511089	DNA
	• AF068548	Mustela vison cytochrome b (cytb) gene, partia...	Neovison vi...	337	AF068548	3273799	DNA

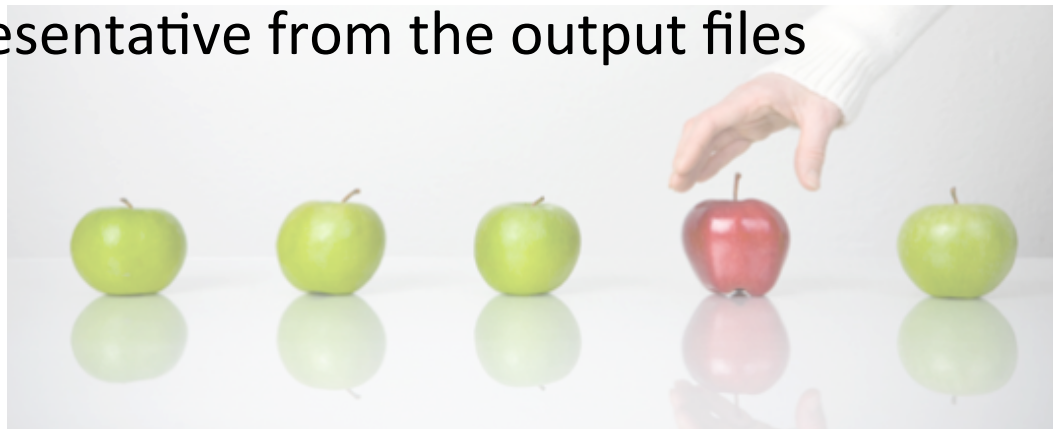
Combine and align with any unpublished data

- Alignment for orientation (5'-3' direction)
- Determine if you need to reduce redundancy
 - All exact matches should be reduced to single sequence



Reducing redundancy

- Probes can anneal to sequences up to 20% (Hawkins et al. 2015 *in revision*) divergent, but better to estimate 10-15% divergence to avoid losing molecules
- Can cluster through CD-HIT-EST to remove overly-similar reads
 - Upload fasta file, determine threshold, use a single representative from the output files



CD-HIT Suite: Biological Sequence Clustering and Comparison

Server home	cd-hit	cd-hit-est	h-cd-hit	h-cd-hit-est	cd-hit-2d	cd-hit-est-2d	result
-----------------------------	------------------------	----------------------------	--------------------------	------------------------------	---------------------------	-------------------------------	------------------------

Sequence file and databases

Load Query Fasta file from your computer: No file chosen

☐ Incorporate annotation info at header line

Sequence Identity Parameters

☒ Sequence identity cut-off

Algorithm Parameters

-r: comparing both strands ☒ No ☐ Yes

-G: use global sequence identity ☐ No ☒ Yes

-g: sequence is clustered to the best cluster that meet the threshold ☐ No ☒ Yes

-b: bandwidth of alignment

Alignment Coverage Parameters

-aL: minimal alignment coverage (fraction) for the longer sequence

-AL: maximum unaligned part (amino acids/bases) for the longer sequence

-aS: minimal alignment coverage (fraction) for the shorter sequence

-AS: maximum unaligned part (amino acids/bases) for the shorter sequence

-s: minimal length similarity (fraction)

-S: maximum length difference in amino acids/bases(-S)

Mail address for job checking

Give your mail address:

Upload aligned or unaligned files

Set identity parameter

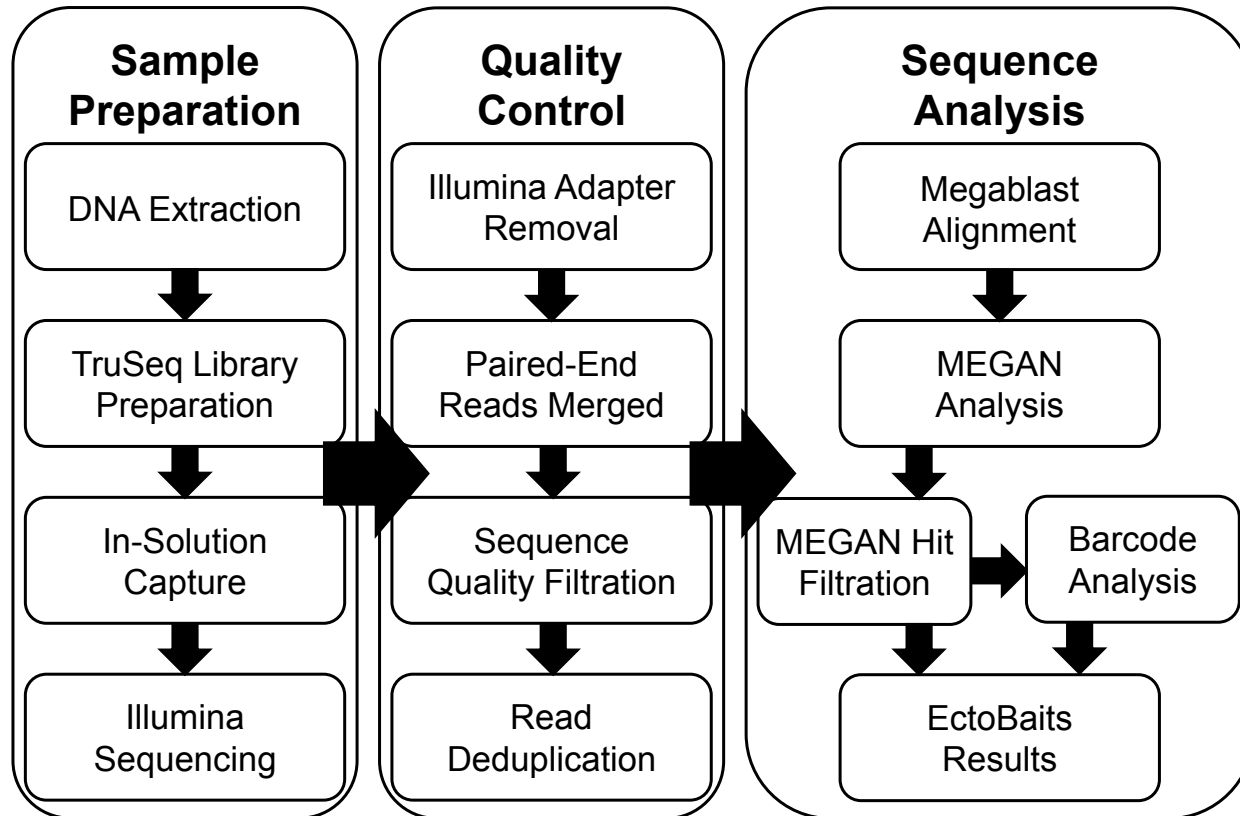
Enter email address

Split reads to probe length

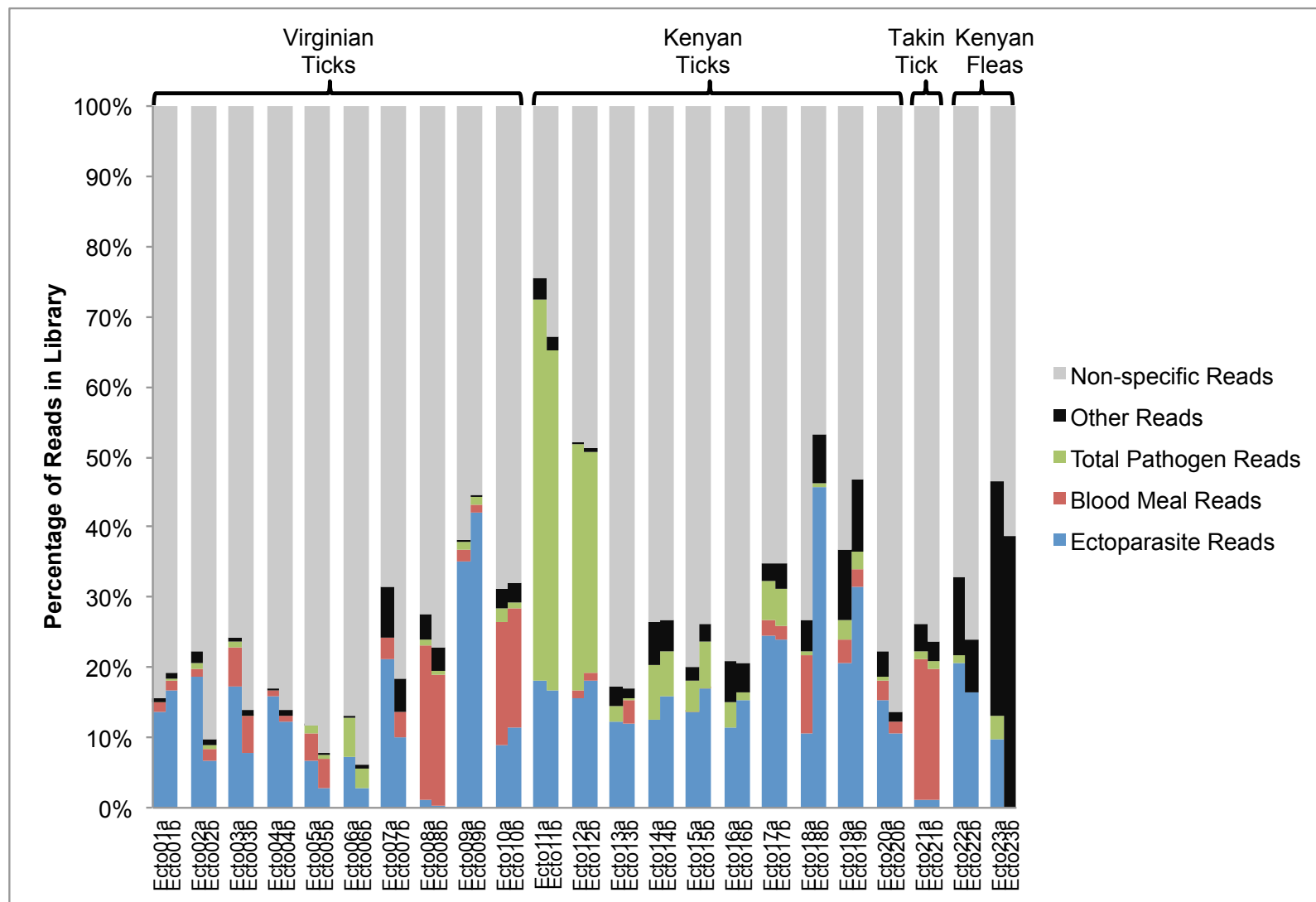
- 'split' command
- EctoBaits: 100bp fragments
 - Mybaits can be from 80-120bp long
- Combine to single fasta file
- Email files to Tech Support
 - Will perform additional QC, replace ambiguous sites
 - Synthesis takes approximately 6-8 weeks

```
1      10      20      30      40      50      60      70      80      90      100
REV ... AAA CGGAGTTAG CCGGTGCTTCTTCTGCGAGTAACGTCAATGATTGAG CGTATTAACTCAA CCCCTT CCT CCT CGCTGAAAGTGCTTTA CAA CCCGAAG
REV ... ATCG CTG CGT CCTT CAT CGTTGTGCGAG CCGAGACATCCACCG CTGAAAATTGATTTTTTAATTATATA CAAAAGTAGTAGTTATATA CAA CAAA CAA
FWD ... TTTTATATGG CCTGAAATTTGATGTGATGATCCGCTGAAATTAAG CATATAA CTAAG CGGAAGAAAAAGAAAAA CAAATGATT CCTTTTCTTTTCTTTT
REV ... AAAGTAATACTATTATATAGTTGACCATATCTAAAAA CAGGCTTACATAATT CAAAACGAAGAATTAAAGATAAA CTAGAT CCTAAAAA C
FWD ... TAGCTGT CAGTTCTACTAAGGTAA CAAAAG CAG CAGATGATG CAGAAAT CG CTTCTG CTCAATTGTTTGATAAAA CAAAGA CATTAGTTGTTT CACTCT
REV ... ATCATG CATAACAGAG CGAA CAT CAG CAAAAT CAAGATTAATGAG CCCTTCTTTAAT CATGAGGT CCGTAATGGAAG CAA CACCCGAATAAAGAACTTGA
FWD ... TATAGGG CA CCAATGATACTGAAGCTACGAATATA CAGA CTATGAAGA CTTAAG CTTTGACT CTTATATAAATTCCAA CAT CAGAATTAAAA CCAGGAGAA
REV ... AAATAAATTTTAATATTTAAATGTGTGTTTCAACATGTGTAAT CAAAAA CTTTATTA CTAAGATGTGAATGAT CTTCCG CAGGTT CA CTTA CCG
REV ... AACG CACGA CCAAAATTTAA CTAATAATTTAA CAAATAAGAAATGATCCTT CCG CAGGTT CA CTTACGGA CTTGTTA CCA CTTCT CTTTCTTTTAA
REV ... AACAAAGAA CTCTGATAAGAAATGAT CTTT CCG CAGGTT CACCTACGGAACCTTGTGA CCA CTTCT CTTTCTTTAAGTATAAGGTT CA CAAA CTTT
REV ... AAAAAAAAAAAAAAAAAAGAT CTTCTG CAGGTT CA CTTA CCGAAA CTTGTTACGACTT CTTCTTCTTTAAGTATAAGGTT CACAAA CTT CCTAGA
FWD ... CCGTAGGTGAACCTGCGGAAGGGTCA TT CA C CTTAGTTTAAATATATATTTTGTGA CAA CAGTAGAAAA CAA CTTTATAA CTATAAAGT CACATTT
REV ... TCAGATTATGAAAAATTTA CTAATA CGG CTGTTAATGAAATGAATGAG CCTATAGAGGAGA CTGTGTTT CATGTGGTATAAG CGT CAGGATAATCAGAG
REV ... AAAATAA CT CTTATTAGAGAGATAATTTGA CCAATTTGATGATAGGAGATTT CATGAGTAAAAAAGT CAGGATAATCTGAATAT CTT CCGGGGCATTCCTCA
REV ... AAATTA CACCCAATAAAGAGATAATTTGAT CCTATTGATGATATGAAATTTCA TTTTATAAAAAAATCTGGGTAGTCTGAATAT CTT CCGAGGTATTTCTCA
REV ... ATTA CGG CTGTTATTGAGATTAAAGAGCCGATGGAGGATATGGGTCTT CATAGGGGTGATG CGTCTGGATAGTCTGAATAT CTT CTTGGTATG CCGG CTA
REV ... CTTGATAAGGAAATAATAGAACCAATTTGATGAAAT CAGATTT CATTTTGAGAAAAAAT CTGGGTAGT CAGAAATAA CTT CTTGG CATT CTTCTTAGA CTTA
REV ... CCAATTAAGGAGATAATTTGAG CTTAGAGAAGAAAGTAAATTT CATTTTGAGAAAAAAGT CTGGGTAACT CCGAATAAT CTT CTTAGGTTG CTTCTTAACCTTA
REV ... CTAATAGAGAAATTAATCTACCTAATGAGGAAATATATTT CATTTAGTAAAAAAT CTGGATAAT CAGAAATA CCGT CAGG CATA CTTCTTAGT CTTA
REV ... ATAGAAGAGATCATTTCTACCAAAAGATGAAATTTAGTTT CATTTAGAGAAGAAAT CCGGGTAACT CAGAGAT CTT CTTGGTATA CTTCTAAGA CCAAGA
REV ... ATAGAAGAAATTTATTTCTCCAGGGAAGAAATTAAGTTT CATTTAGAAAAGAAAT CAGGATAAT CTGAATAT CTT CTTGGGTAT CTTCTTTAAT CTTAAAA
REV ... ACTG CAGAAATTTCTT CCTTAATGAGGAAATCTAAATTT CATTTAGAGAAGAAAT CAGGAAAT CAGAAATA CTT CTTGGGTAT CTTCTTTAAT CTTAAAA
```

EctoBaits Analysis:



EctoBaits Results:



Other Capture Array Designs:

- Mammal mitogenome array 'MMA'
- Bacterial Genomes (homologous regions)
- Ultraconserved Elements
- RAD-seq SNP ID'd contigs
- Limited only by creativity



Questions?

