

DISCOVERY AND CHARACTERIZATION OF NOVEL SIGNATURES FROM THE *RICINUS COMMUNIS* (CASTOR BEAN) GENOME

Kevin P. O'Connell* and Evan W. Skowronski, Research and Technology Directorate
US Army Edgewood Chemical Biological Center, Aberdeen Proving Ground, Maryland 21010

Kenneth L. Dretchen and Jonathan A. Leshin, Department of Pharmacology,
Georgetown University, Washington, DC 20057

Andrea Weeks, Department of Environmental Science & Policy
George Mason University, Fairfax, Virginia 22030-4444

ABSTRACT

Given the infamous toxicity of ricin and the industrial usefulness of castor oil, there is a surprising lack of information about the genetic diversity of the species that produces both, the castor bean plant (*Ricinus communis* L.). The extent of DNA sequence variation in the gene that encodes ricin (preproricin) is also poorly understood. This lack of knowledge hampers the ability to make effective assays, or to associate ricin toxin from sites of release with suspect production labs. Without this basic genetic understanding, assays for ricin may not positively react with genomic DNA from *R. communis* derived from any source. We are remedying this shortfall by pursuing a genetic characterization of members of *R. communis* collected from around the world. Preliminary data from the amplification and sequencing of preproricin genes from 63 members of this collection indicate the presence of a large number of nucleotide polymorphisms, and the possible presence in some varieties of a previously unreported, shorter-length paralog of the preproricin gene.

1. INTRODUCTION

Ricin is a highly toxic protein found in the seeds of *R. communis*. The toxin is listed as a CDC Select Agent and is controlled under DoD surety regulations. The extraction of castor oil overseas results in the annual production of thousands of tons of castor bean mash, which contains between one and ten percent ricin by weight. The attraction of terrorists and criminals to ricin is apparent from its discovery in a makeshift lab in London in 2003, its presence in a letter mailed to the US Senate majority leader in 2004, and its discovery with bomb-making materials in a suburban home in Tennessee in mid-2006. Ricin was likely the poison used by an assassin to kill Gyorgi Markov, a Bulgarian dissident living in London in the late 1970's. The toxin was contained in a pellet injected into Markov's leg by a device disguised as an umbrella.

Mature ricin has two subunits (A and B chains) that are derived from a single protein (preproricin) encoded by one or more genes in the *R. communis* genome. During

maturation of the toxin protein, a leader peptide before the A chain sequence and a linker peptide between the A and B chains are removed by proteases. In nature, ricin likely serves as a defense against seed-eaters by inactivating ribosomes and stopping protein synthesis, an activity that renders it toxic to animal cells. It does so by catalytically depurinating a portion of the large subunit ribosomal RNA (Lord *et al.* 1994). The catalytic nature of the endonuclease portion of the toxin (the A subunit) is the source of the extreme toxicity of the protein; theoretically, a single ricin molecule can kill a cell by inactivating all ribosomes present (Wiley and Oeltmann 1991). Reported estimates of the mouse LD₅₀ (by inhalation) range from 0.1 µg/kg body weight (Whalley 1990; Parker *et al.* 1996) to 3-5 µg/kg (Kortepeter *et al.* 2001), corresponding to 7 to 350 µg for a 70 kg adult human (depending on the purity of the toxin). Domestically, the plants grow wild, and seeds can be purchased commercially without restriction.

Although *R. communis* is thought to have evolved in NE Africa (Carter and Smith 1987), populations of this species are distributed world-wide due to its weedy habit combined with its value as an oil-crop and ornamental plant. In frost-free environments, *Ricinus* persists as a large (3 m) woody perennial in disturbed habitats. In agricultural settings or in temperate climates, *Ricinus* survives only one growing season and is propagated from year to year by seed.

R. communis comprises 22 subspecies and varieties as well as a handful of cultivars developed by ornamental horticulturalists and plant breeders. It is the only species in its genus as well as its subtribe, Ricinae, in the tribe Acalypheae, subfamily Acalyphoideae of the family Euphorbiaceae (Webster 1994). Molecular phylogenetic analyses indicate *Ricinus* is mostly closely related to *Sperkansia*, a small genus native to China, and confirm that *Ricinus* is part of a natural lineage containing those genera in the tribe Acalypheae (Wurdack *et al.*, 2005). However, measures of infrappecific genetic diversity either within the genus as a whole, or among the wild varieties and cultivars are completely lacking.

Report Documentation Page				Form Approved OMB No. 0704-0188	
Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.					
1. REPORT DATE 01 NOV 2006		2. REPORT TYPE N/A		3. DATES COVERED -	
4. TITLE AND SUBTITLE Discovery And Characterization Of Novel Signatures From The Ricinus Communis (Castor Bean) Genome				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S)				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) US Army Edgewood Chemical Biological Center, Aberdeen Proving Ground, Maryland 21010				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release, distribution unlimited					
13. SUPPLEMENTARY NOTES See also ADM002075., The original document contains color images.					
14. ABSTRACT					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT UU	18. NUMBER OF PAGES 29	19a. NAME OF RESPONSIBLE PERSON
a. REPORT unclassified	b. ABSTRACT unclassified	c. THIS PAGE unclassified			

An extensive literature and GenBank sequence survey reveals that there is little published information regarding the molecular diversity of the genes that encode ricin that could assist a forensics investigation. Such data, in effect “fingerprints” of ricin genes, and knowledge about the overall genetic diversity of *Ricinus communis* varieties worldwide, are required to establish assays that can identify and characterize a ricin sample regardless of its genetic or geographical source. In a completely analogous effort, the existence of such data for *Bacillus anthracis* and related bacteria allowed the rapid identification of the anthrax strain used in the 2001 biological attacks on the U. S. Senate and media. Such detailed analyses for (prepro)ricin and *R. communis* are currently not possible.

The value of basic genetic data in the analysis and detection of ricin cannot be overstated. Unless potential adversaries invest significantly in expensive, painstaking purification of ricin, ricin extracted from castor beans will be crude and the resultant ricin concentrate will contain impurities, including whole genomic DNA from the *R. communis* seeds. Laboratory tests to determine whether a particular sample contains ricin can depend on the ability of existing primers to amplify *Ricinus* DNA (the preproricin gene) using PCR. It is also now unknown whether PCR assays for ricin-encoding genes will positively identify all ricin gene samples, because only four native ricin gene sequences are available in GenBank, the public database of gene sequences managed by the National Center for Biotechnology Information. Knowledge of the natural diversity in ricin gene sequences is required to design effective and universally applicable assays for this agent.

2. MATERIALS AND METHODS

Germplasm. Samples of *R. communis* represent the countries listed in Table 1. Materials were obtained from USDA, and commercial and academic sources. Seeds were planted in Miracle-Gro potting mix and grown under natural daylight in a greenhouse at ECBC.

DNA purification and amplification. DNA was prepared from leaf tissue ground in a mortar and pestle under liquid nitrogen, using Qiagen DNeasy Plant Maxi kits according to the manufacturer’s instructions, and quantified spectrophotometrically. Primers were designed to amplify the preproricin gene from *R. communis* genomic DNA using the LaserGene software suite (DNASTAR, Madison, WI) and GenBank accession X02388 (preproricin). Primers were synthesized commercially. PCR reactions contained 45µl of PCR Supermix (Life Technologies, Gaithersburg, MD), 1µl of each primer (20 pmol), and 3µl of genomic DNA for a final reaction volume of 50 µl. PCR reactions were

performed on an Applied Biosystems 9600 thermocycler using the following temperature cycling profile: 94°C / 5 min; 30 cycles of 94°C / 1 min, 55°C / 1 min, 72°C / 1 min; 72°C / 5 min, followed by a 4°C soak. The products of each reaction were separated by electrophoresis on 0.8% agarose gels and visualized by staining with ethidium bromide. All products were cloned using Topo-TA kits (Invitrogen) and verified via reamplification from positive colonies using plasmid primers. Three clones of each amplicon were randomly chosen for sequencing.

Sequencing. Sequencing reactions were set up according to the reagent manufacturer’s recommendations. Briefly, 10-50 ng of PCR product or 100-500 ng of plasmid DNA was added to 1/8X ABI Big Dye Terminator v3.1 Cycle Sequencing reaction (Applied Biosystems Inc., Foster City, CA) and run with the standard cycling protocol. Individual sequencing reactions were performed with two pairs of nested internal primers for each PCR product/plasmid in order to give high quality, overlapping sequence of the entire amplicon. Sequence reaction cleanup was performed with DyeEx 2.0 terminator removal kits (Qiagen, Valencia, CA). After electrophoresis on an ABI 3130, automated sequence base calling and quality scoring was performed with the ABI sequencing software.

Table 1. Countries currently represented by the ECBC castor bean sample collection.

<u>Africa</u>	<u>Americas</u>
Cameroun	Argentina
Egypt	Bolivia
Ethiopia	Ecuador
Kenya	Mexico
Malawi	Peru
Nigeria	USA
South Africa	
Sudan	
Uganda	
<u>Central/Middle East</u>	<u>South East Asia</u>
Afghanistan	Hong Kong
Former Soviet Union	India
Iran	Indonesia
Israel	Japan
Jordan	Malaya
Pakistan	Philippines
Saudi Arabia	Sri Lanka
Syria	Taiwan
Turkey	Vietnam

Individual sequencing reads for a given clone were assembled into a single consensus sequence using DNA analysis software, and verified by an experienced operator. Ambiguous or missing sequence were repeated and added to the assembly as needed to obtain overlapping, high quality sequence.

3. RESULTS

3.1 Preproricin amplification.

As part of our overall strategy to characterize the diversity of preproricin genes, we obtained samples of castor bean varieties originating in several countries (Table 1) including the former Soviet Union, Iran, Pakistan, Turkey, India, Syria, Israel, Afghanistan, Kenya, and the US, from academic, government and commercial sources. We designed PCR primers to amplify nearly the entire preproricin gene, encoding 512 amino acids of the preproricin protein.

R. communis has been reported to contain as many as eight preproricin genes or pseudogenes (by Southern hybridization analysis); it is possible that the “extra” bands observed in Southern blots correspond to smaller variants or mutants of preproricin, or related proteins. Beyond their observation in Southern blots, no information identifying the origin of these sequences was reported. Our preliminary data support this hypothesis.

Multiple bands were observed following agarose electrophoresis of PCR amplification reactions using the preproricin primers described above (Fig. 1) of approximate sizes 1700 bp, 1400 bp and 1200 bp. The largest bands represent molecules that we predicted by size to encode nearly the entire gene. The lower molecular weight bands, where they occurred, were consistent in size among the cultivars from which they were amplified.

3.2 Preproricin gene characteristics.

Each of the 63 *Ricinus* plants yielded one to three differently sized DNA molecules with the preproricin amplification primers, the largest of which was present in all plants and corresponded to known, functional copies of the preproricin gene in size. We sequenced 126 of the large bands amplified from 63 *R. communis* varieties and compared the sequences to each other and those published in GenBank. At the nucleotide level, differences in preproricin sequence between pairs varied from 0 to 41 bases (mean 7.4 ± 4.9 bp) with a transition / transversion ratio of 0 - 9 (mean 2.2 ± 1.7 bp). The percentage of adenine/thymine bases was 57.6%. Among all 126 sequences and across all 512 AA of the alignment, there were 190 predicted nonsynonymous amino acid (AA)

changes out of 64,512 possibilities (126 sequences x 512 AA per sequence) (Fig. 2).

The amplification of preproricin DNA from some castor plant varieties produced one or two additional, smaller bands. Sequencing 50 clones of the smallest of the three bands revealed that these amplicons, derived from 16 different castor varieties, all shared a single deletion (Fig. 2). The small amplicon is missing base pairs 724-1058 (representing the last 12 codons of the A-chain end, the entire linker peptide, and the first 77 codons of the B-chain). We hypothesize that the small amplicon arises from an altered, duplicate copy of the preproricin gene (a paralog) that is shared among the 16 different castor plants because of common ancestry. If so, this is the first detailed description of such a deletion in the preproricin gene and an important advance in our understanding of the preproricin gene family. The first 242 AA of this paralog are in frame, which suggests that, if actually expressed, it may yield functional transcripts *in vivo*.

The single nucleotide polymorphisms (SNPs) and non-synonymous amino acid changes appeared to be distributed along the length of the large molecule, including one in the linker portion that joins A and B subunits (Fig. 3). The presence of a number of SNPs in the preproricin gene suggests that there is sufficient sequence heterogeneity to present a challenge to workers attempting to construct nucleic acid-based assays to detect preproricin sequences. The heterogeneity in preproricin sequences is also evidence that such differences will indeed yield markers useful in distinguishing plants by preproricin genotype.

3.3 Phylogenetic data analysis.

Although pairwise nucleotide variation is high among sequences of preproricin, the percentage variation across the whole dataset is too low to resolve any particular phylogenetic topology with credible statistical support. Parsimony analysis of the data indicates the matrix of 126 sequences provides only 54 phylogenetic informative characters (3.5% of 1537), too few to yield anything other than few resolved nodes (Fig. 4) but typical of gene sequence data at the population level. Distance analysis of the same data using neighbor-joining yields a more highly resolved tree (data not shown) as a consequence of the algorithm but has virtually no bootstrap support for any of the groupings and extremely small divergences between castor varieties. Consequently, preliminary data suggest that we cannot rely on phylogenetic analysis of preproricin sequences to tell us much about the relationships between varieties of castor bean plants or their geographic origin. This result is expected because DNA sequences of single genes generally do not provide adequate phylogenetic resolution

for population-level studies of plants (Avisé 2004), such as our broad survey of preproricin genes in *Ricinus*.

Of the 1537 bases in each of the 126 preproricin sequences studied, 1262 bases (82.1%) are completely invariable. However, another 221 bases (14.4%) occur in only one of the 126 sequences. While we know that some sequences are absolutely identical with each other, on average this means that each preproricin sequence in our dataset could be uniquely distinguished with 1.8 (221 / 126) SNPs. *Therefore, fingerprinting studies for forensic attribution cases might want to focus on understanding SNP variation within preproricin sequences, in addition to searching the whole genome with other methods.*

Cloned preproricin sequences from individual castor bean plants routinely contain unique genetic signatures. It is unknown whether preproricin sequence variants within a single castor bean plant are a consequence of biochemical artifacts, human editing errors, or true genetic variation. If they do record true genetic variation within the genome of a single castor bean plant as we hypothesize, these sequences may be alleles from one or more loci. The fact that the three-clone “samples” from individual castor bean plants repeatedly yield three different preproricin sequences suggests either that error is pervasive or that there are two or more loci of functional preproricin genes, as one would expect only 2 alleles per loci in a diploid organism. Biochemical error is unlikely because basepair error would be unbiased towards maintaining a reading frame, and the preproricin sequences observed were always in frame. Few human editing errors (less than 10 instances) were identified prior to data analysis on the basis of the presence of anomalous

stop codons in the matrix and were corrected after referring to the original, unedited sequence chromatograms.

Therefore, evidence points most strongly towards two or more functional preproricin loci in the castor bean genome. This indicates that further investigation of SNP variation at the allelic level among different castor bean plants may yield more extensive variation than uncovered by our preliminary analyses, because the number of potentially different alleles doubles with each additional locus discovered. *Additional sequence characterization of the preproricin gene family is likely to reveal that SNP analysis of preproricin gene sequences is a statistically powerful tool for forensic attribution.*

4. CONCLUSIONS

These data are beginning to define the genetic markers that may allow identification of ricin toxin from the broadest variety of *R. communis* studied so far. The survey of ricin gene sequences will have applications in several areas of basic and applied research:

- The three-dimensional structure of ricin has been resolved by x-ray crystallography (Marsden *et al.* 2004). Because the presence of certain amino acids at given positions in the protein is correlated with toxin activity, preproricin sequence data obtained from the proposed study, when combined with the emerging understanding of ricin structure/function relationships, may enable us to predict the toxicity of ricin preparations.



Figure 1. Amplification of *R. communis* genomic DNA using primers for the preproricin gene. Note the presence of multiple bands in some lanes. Each variety is identified by its USDA plant introduction (PI) number. Lane 1, 1-kilobase ladder (DNA size markers); lane 2, PI167112; lane 3, 167287; lane 4, 167342; lane 5, 170684; lane 6, 170686; lane 7, 222265 1407; lane 8, 222745; lane 9, 222828; lane 10, 222829; lane 11, 222830; lane 12, 223408; lane 13, 223409; lane 14, 227869; lane 15, 229540; lane 16, 229541; lane 17, 229620; lane 18, 229785; lane 19, 250880; lane 20, 250881; lane 21, 250884; lane 22, 250938; lane 23, 250942; lane 24, positive control; lane 25, negative control.

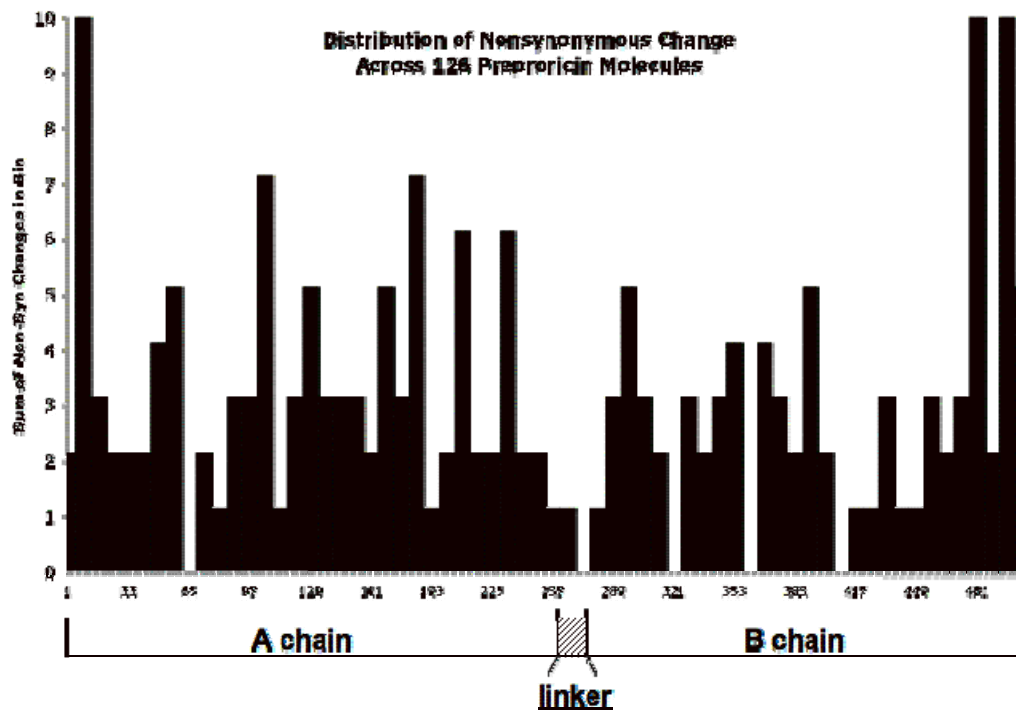


Figure 2. Distribution of non-synonymous changes along the length of the preproricin molecule, from a sample of 126 sequences derived from 63 different castor bean plants. The original 1715 bp fragments were trimmed to 1537 bp (512 amino acids plus one base) to exclude missing data among some sequences in the alignment.

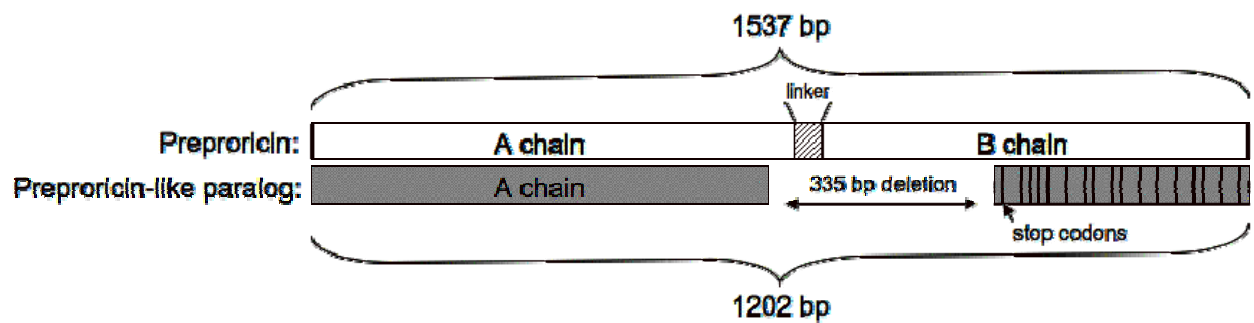


Figure 3. Comparative nucleotide alignment of the small amplicon (preproricin-like paralog) and the large amplicon (preproricin).

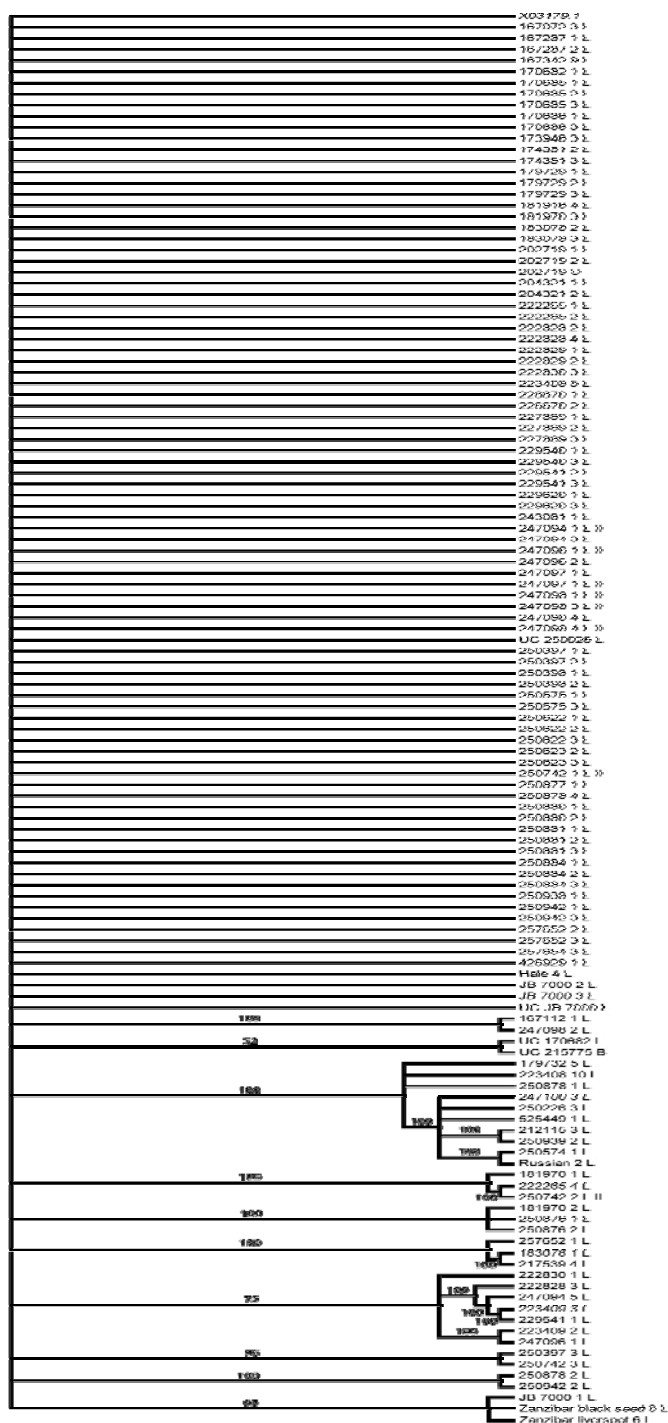


Figure 3. Lack of significant phylogenetic information in full-length preprorin gene sequences. Displayed is the 50% majority rule of 80,000 most-parsimonious trees. A similar analysis of the sequences of the shorter preprorin paralog also revealed that the sequence data revealed no statistically significant phylogenetic differences among the plants tested (data not shown).

- The same knowledge base will also help predict both reactivity versus antibodies employed in detection systems (such as hand-held assays), and the toxicity of ricin preparations. The latter also help determine whether current antibody detection methods work against the world-wide allelic diversity of ricin toxin molecules and whether ricin is produced by other genera related to *Ricinus*.

- Reagentless methods will also benefit from the preprorin sequence data, because the primary sequence allows the prediction of the molecular weight of the toxin, permitting the identification of ricin varieties by methods such as mass spectrometry (MS) more accurately than is now possible. Positive identification of proteins by MS requires prior knowledge of the mass of target proteins; the mass of ricin molecules produced by each plant variety can be predicted to the Dalton using DNA sequence to predict protein primary structure.

By obtaining the largest number of ricin gene sequences and other common *R. communis* signature sequences assembled to date, we will gain the ability to determine whether current gene probe assays to detect ricin are effective against the global diversity of ricin alleles, provide new gene targets, and populate new protein toxin mass reference libraries for MS.

Potential non-CB defense spin-offs may include:

- The discovery of ricin toxin variants with altered pharmaceutical properties. Ricin A chain has been studied in combination with tumor-specific (tumor-targeting) antibodies for the construction of so-called immunotoxins, toxins that preferentially kill tumor cells in cancer patients. The identification from sequence data of ricin with a variety of activities may facilitate the development of novel chemotherapeutics.
- Plant breeding markers: the amplified fragment length polymorphism (AFLP) data to be developed in this study (see below) will provide the basis for constructing a genetic map for castor plants that may assist in the breeding of *R. communis* for improved oilseed production (castor oil is also a raw material in the National Strategic Stockpile). Likewise, AFLP markers and their mapping relative to preprorin genes will assist in the breeding of ricin-free castor plants, a “holy grail” of castor researchers and agronomists who seek to revive the US castor oil industry.

- Forensics: Like the sequence data described above, AFLP fingerprinting data may also be useful in the forensic association of weaponized material with a geographic origin for the source plants, or for associating ricin preparations with suspect manufacturing facilities. Our project leverages a draft genome of the US cultivar Hale, recently submitted to GenBank as a result of project sponsored by the US Department of Justice. Such data are useful as a baseline against which to compare the diverse sequences being generated by our study.

We are continuing to characterize the preproricin gene and to extend our understanding of the intraspecific genetic diversity of *R. communis* by identifying and analyzing additional genetic markers (AFLP, whole organellar sequences) for evolutionary significance and that may be useful as signatures for distinguishing varieties or individual plants.

REFERENCES

- Avise, J. C. 2004. Molecular markers, natural history, and evolution. 2nd ed. Sinauer Associates, Inc. Sunderland, MA.
- Carter, S. and A. R. Smith, 1987. Euphorbiaceae. In: Flora of Tropical East Africa. A.A. Balkema Publishers, Rotterdam.
- Kortepeter, M., Christopher, G., Cieslak, T., *et al.*, 2001. *Medical Management of Biological Casualties Handbook*. 4th ed. Frederick, MD: USAMRIID.
- Marsden, C. J., Fulop, V., Day, P. J., and Lord JM. 2004. The effect of mutations surrounding and within the active site on the catalytic activity of ricin A chain. *Eur. J. Biochem.* 271:153-162.
- Lord, M. J., Roberts, L. M., and Robertus, J. D., 1994. Ricin: structure, mode of action, and some current applications. *FASEB J.* 8: 201-208.
- Parker, D. T., Parker, A.C., and Ramachandran, C. K. 1996. Joint CB Technical Data Source Book. Vol. IX. Toxin Agents. DPG document no. DPG/JCP-96/007.
- Tregear, J. W., and Roberts, L. M. 1992. The lectin gene family of *Ricinus communis*: Cloning of a functional ricin gene and three lectin pseudogenes. *Plant Mol. Biol.* 18: 515-525.
- Webster, G. L., 1994. Synopsis of the genera and suprageneric taxa of Euphorbiaceae. *Ann. Missouri Bot. Garden* 81: 33-144.
- Whalley, C. E., 1990. Toxins of Biological Origin. CRDEC-SP-021, Chemical Research, Development and Engineering Center. AD B145632.
- Wiley, R. G., and Oeltmann, T. N., 1991. Ricin and Related Plant Toxins: Mechanisms of Action and Neurobiological Applications. In: Handbook of Natural Toxins, Vol.6. R.F.Keeler and A.T.Tu (eds.) Marcel Dekker, Inc., New York.
- Wurdack, K. J, Hoffman, P., and Chase, M. W., 2005. Molecular phylogenetic analysis of uniovulate Euphorbiaceae (Euphorbiaceae sensu stricto) using plastid *rbcL* and *trnL-F* DNA sequences. *Am. J. Bot.* 92: 1397-1420.

UNCLASSIFIED

Discovery and Characterization of Novel Signatures from the *Ricinus communis* L. (Castor Bean) Genome

Kevin P. O'Connell, Ph.D.

US Army Edgewood Chemical Biological Center

28 November 2006



UNCLASSIFIED

Background: the plant

- Ricin is derived from castor beans, the seeds of *Ricinus communis*.
- Cultivated since antiquity.
- Global production of beans exceeds 1M tons/yr for oil production (30-50% of bean mass).
- After oil extraction, the remaining mash is 1-10% ricin (300,000 tons/yr x 1% = 3000 tons annually)



R. communis



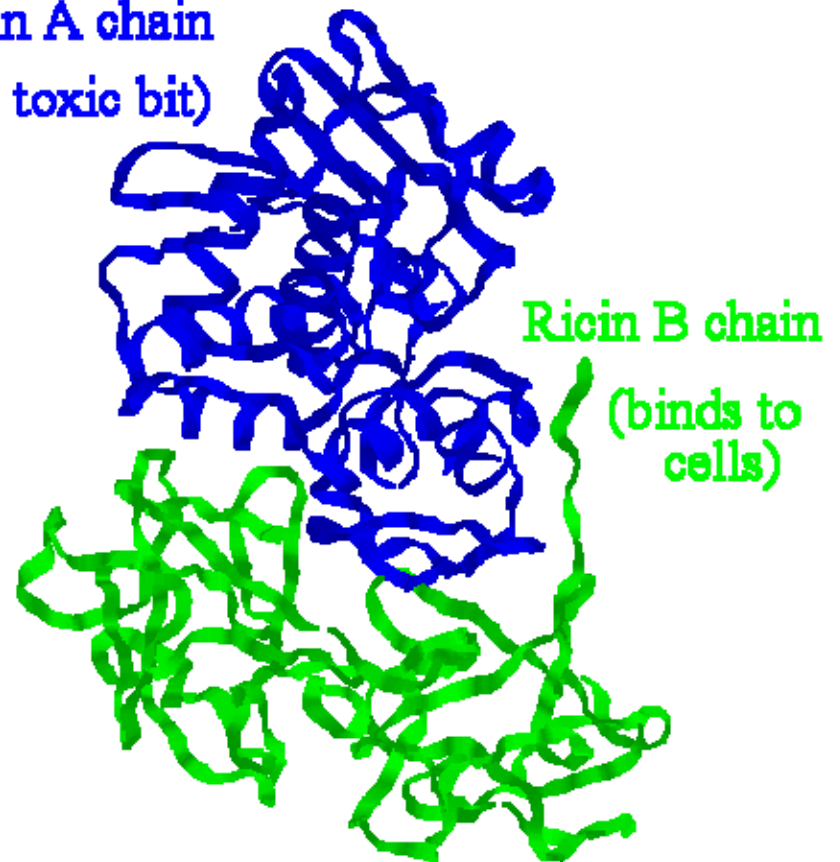
Castor beans



Background: the toxin

- The mature toxin has two subunits, each with its own function.
- B chain is a galactose-specific lectin (a protein that tightly binds sugars) and facilitates entry.
- A chain depurinates a specific base in the 28S subunit of ribosomes (A4324 in mammals), killing the cell.

Ricin A chain
(the toxic bit)

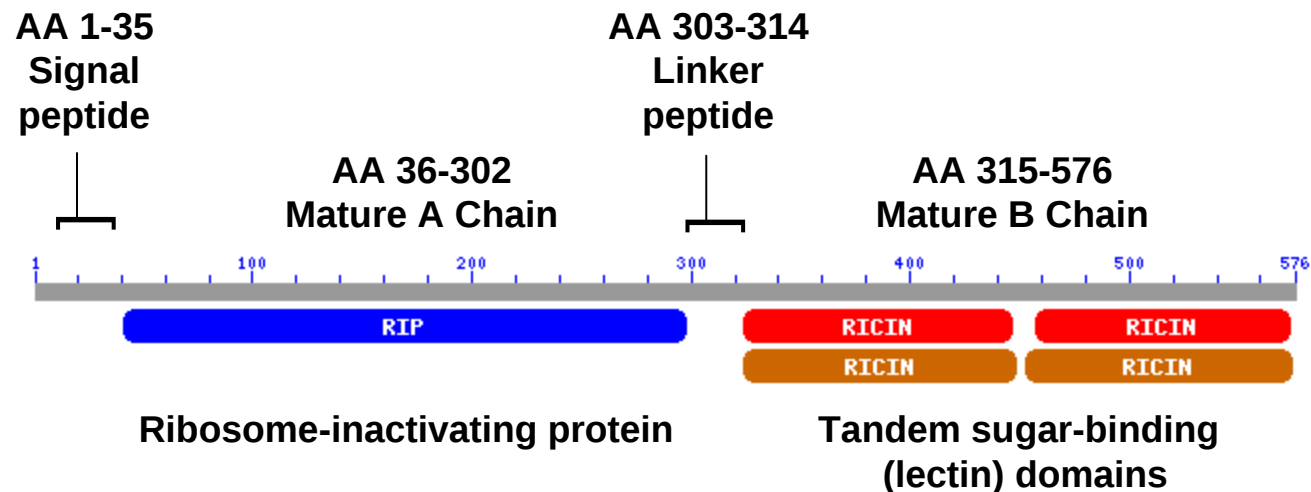


U of Warwick



Background: the gene

- The ricin gene encodes a prepro-toxin of 576 AA.
- A signal peptide precedes the A chain sequence.
- The signal peptide and a spacer peptide between A and B chain are removed during maturation.



UNCLASSIFIED

Detecting ricin:

- Immunological methods (HHAs)
- Mass spectrometry
- Detection of *R. communis* sequences

All three of these methods are impacted by the following Problematic Observations:

- Few sequences (>10) in GenBank for preproricin
- No published work on *R. communis* diversity

Poor understanding of the diversity of the target → uncertainty in assay design and interpretation



ECBC

UNCLASSIFIED

UNCLASSIFIED

Questions: “Novel signatures = diversity”

Basic science:

- What is the extent of natural sequence heterogeneity in the preproricin gene? (DNA sequence analysis, most of this talk)
- What is the copy number of preproricin genes in the *R. communis* genome? (hybridization analysis)
- What is the diversity of *R. communis* as a whole? (AFLP)

Applications in biological defense:

- Implications of sequence heterogeneity for:
 - toxin detection (genetic, immunological, physical methods)
 - threat assessment (variants of differing toxicity?)

Other applications:

- Forensics (attribution), pharmacology (anti-cancer immunotoxins), oilseed production (strategic material), crop reintroduction (grows on marginal land)

UNCLASSIFIED



ECBC

UNCLASSIFIED

Preproricin heterogeneity: approach

Obtained over 95 seed samples from USDA, commercial sources

Major sources: Iran (38), Turkey (11), Pakistan (10), India/Sri Lanka (7), FSU (4), Syria, Afghanistan, Egypt (2 each)

Grew plants to obtain leaf tissue for DNA extraction

Designed primers to amplify preproricin
(primer design constraints)

Purified amplicons by agarose gel
electrophoresis, cloned amplicons
into vectors for sequencing



UNCLASSIFIED



ECBC

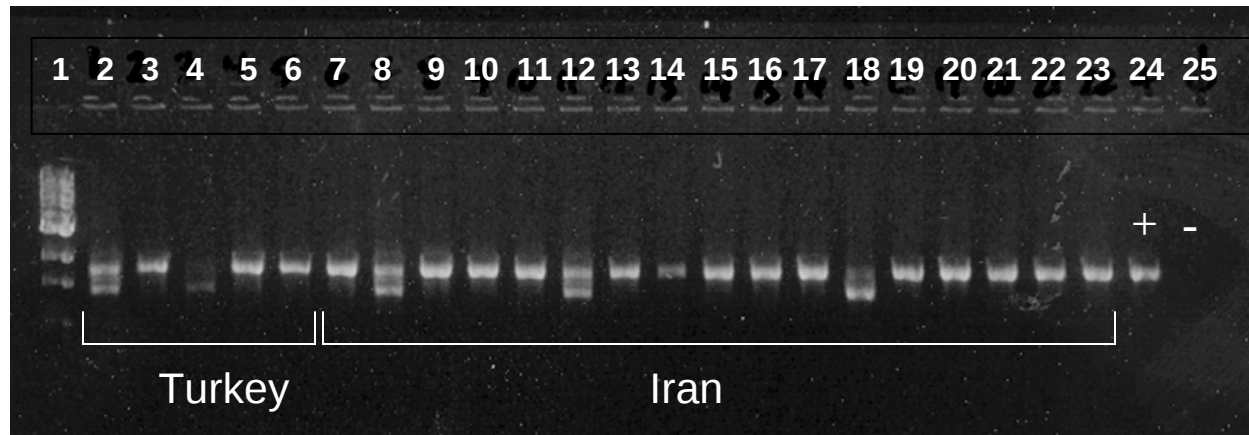
UNCLASSIFIED

Preproricin heterogeneity: results

Unexpectedly, gel analysis of amplicons revealed that many plants yielded 2 or 3 bands: ~ 1700 bp (“full length”), 1600 bp, 1400 bp

Multiple clones obtained from most bands from 63 accessions.

(Aside: pressed herbarium specimens prepared from sequenced plants.)



Lane 2, PI167112; lane 3, 167287; lane 4, 167342; lane 5, 170684; lane 6, 170686; lane 7, 222265 1407; lane 8, 222745; lane 9, 222828; lane 10, 222829; lane 11, 222830; lane 12, 223408; lane 13, 223409; lane 14, 227869; lane 15, 229540; lane 16, 229541; lane 17, 229620; lane 18, 229785; lane 19, 250880; lane 20, 250881; lane 21, 250884; lane 22, 250938; lane 23, 250942

UNCLASSIFIED



ECBC

UNCLASSIFIED

Preproricin heterogeneity: sequence analysis

Sequenced both strands of 126 “large” amplicons from 63 plants

Contigs trimmed to 1537 bp to remove ambiguous calls; alignment is missing 3 N-terminal AAs from A chain and 28 C-terminal AAs from B chain (yielded 512 AAs when proteins were aligned).

No indels relative to sequence reported by Halling (1985), and no stop codons.

HOWEVER: most sequences are unique (114 haplotypes). Unique in what way?

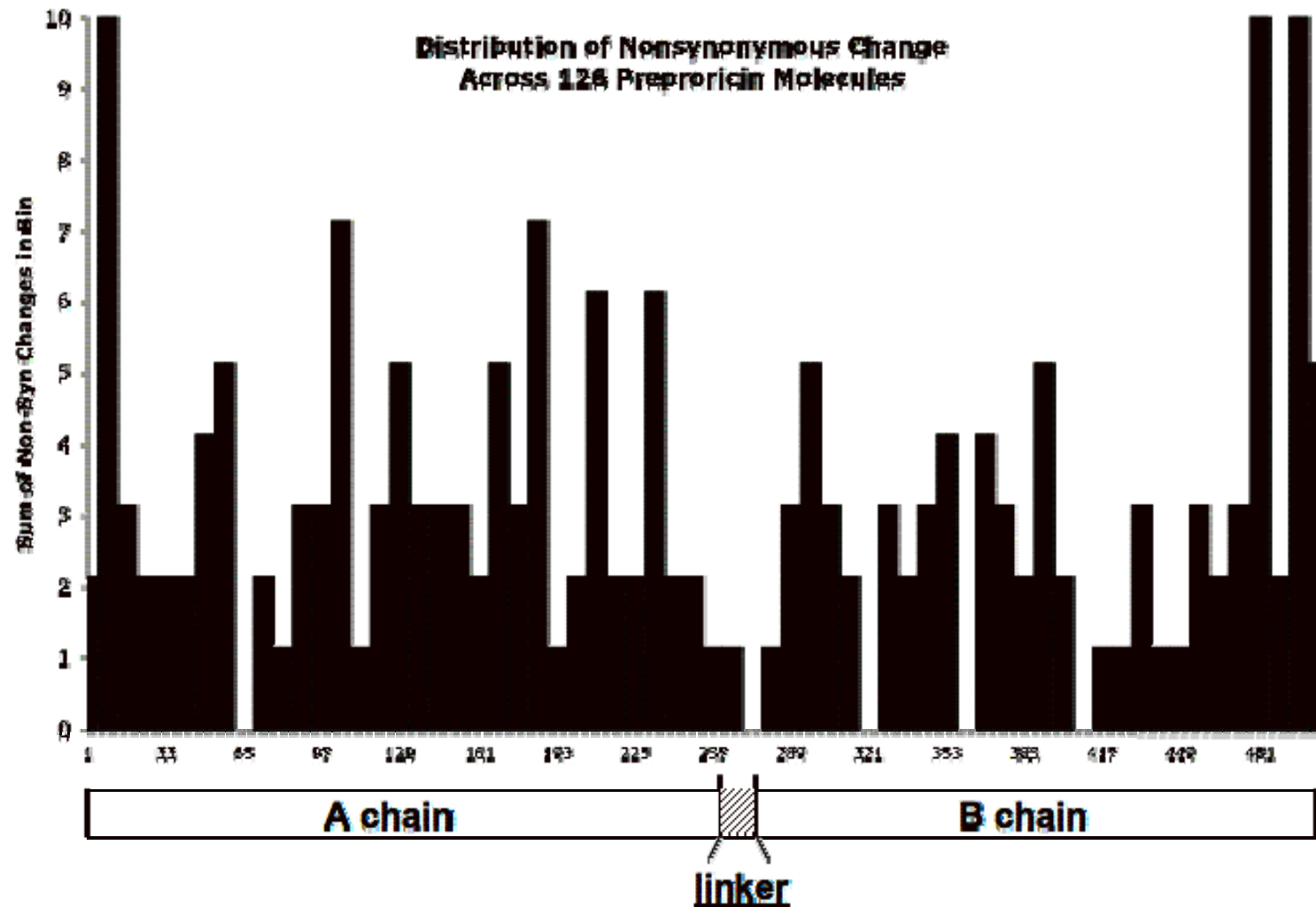
- 190 nonsynonymous changes out of 126 X 512 possibilities, distributed along the length of the molecule



ECBC

UNCLASSIFIED

UNCLASSIFIED



126 preproricin sequences, including 122 sequences derived from clones and four sequences derived from uncloned PCR product. In total these sequences represent 63 different castor bean plants.

UNCLASSIFIED



ECBC

Preproricin heterogeneity: sequence analysis

Possible reasons for the observed heterogeneity:

1) Sequencing error?

Not likely: random error would introduce stop codons at many positions, and we see none

2) Multiple copies of the gene? True.

R. communis is diploid

We and others detect more copies by Southern hybridization

3) Evolutionary drift? Possibly.

Few changes in AAs known to have direct roles in A or B chain functions

Others AAs have more freedom to drift

A chain: no changes (relative to Halling) in Ala165, Arg180, Arg196, Trp211, Leu214, Pro229, Arg235. Only 6 sequences had nonsynonymous changes in critical AAs

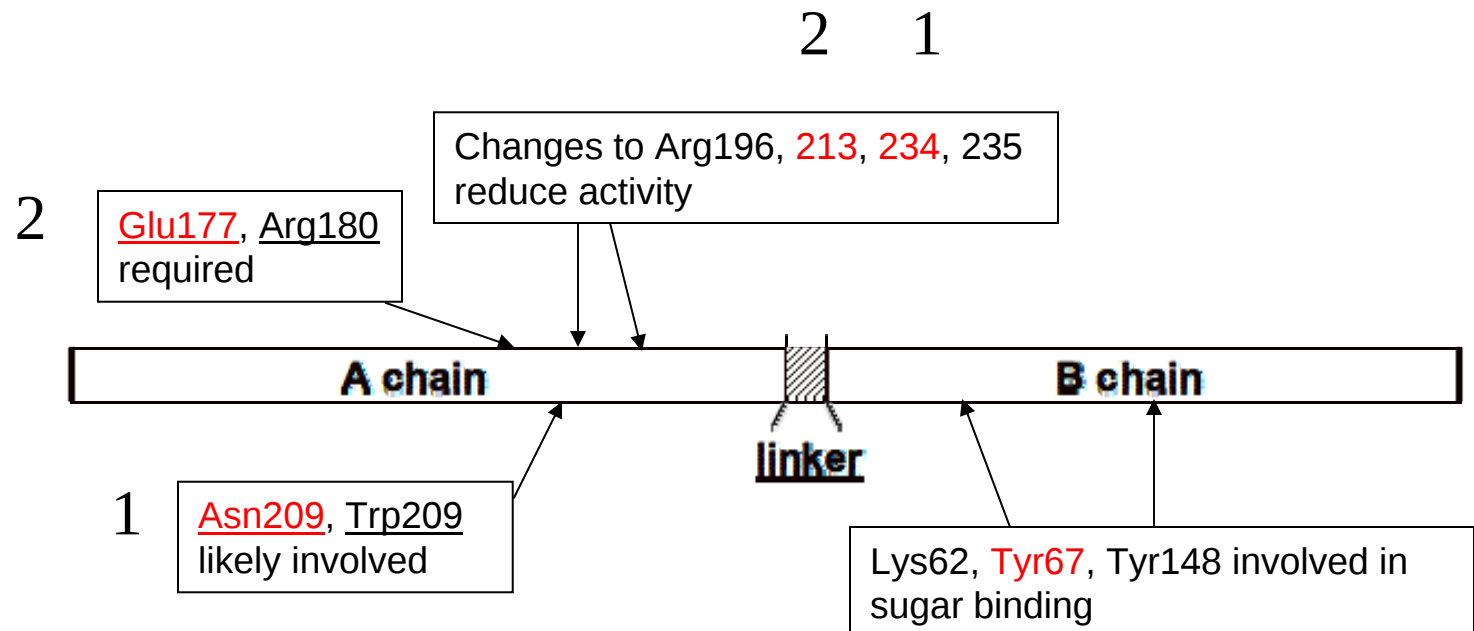
B chain: no change in Lys62 or Tyr148. One conservative change in one sequence (Tyr67→Phe).



UNCLASSIFIED

A chain: no changes (relative to Halling) in Ala165, Arg180, Arg196, Trp211, Leu214, Pro229, Arg235. Only 6 sequences had nonsynonymous changes in critical AAs (underlined AAs conserved among RIPs)

B chain: no change in Lys62 or Tyr148. One conservative change in one sequence (Tyr67→Phe).



UNCLASSIFIED



ECBC

UNCLASSIFIED

Preproricin heterogeneity: Phylogenetic analysis

Most SNPs observed are unique to a particular clone.

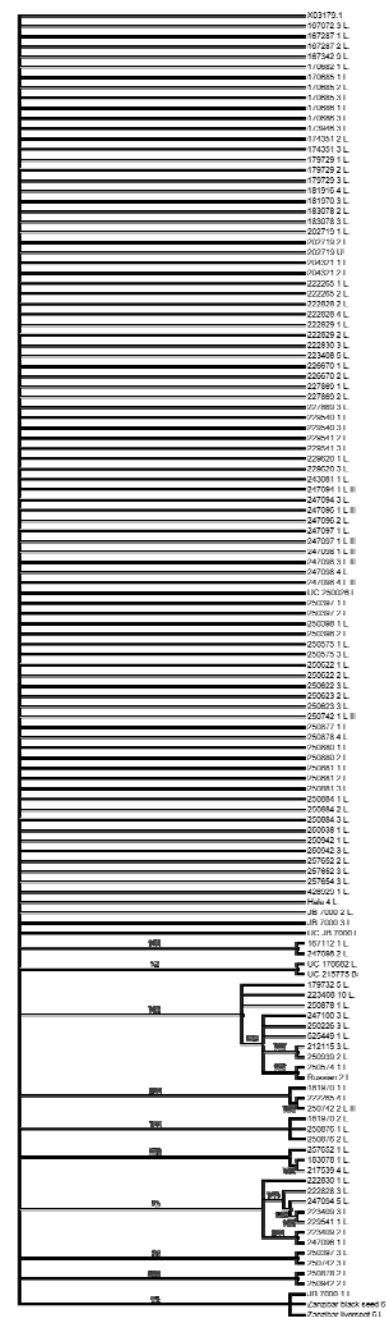
Such SNPs are uninformative in parsimony analysis.

No data from USDA catalog on collection sites within countries. Correlation with country of origin, if possible at all, will depend on more sampling and better knowledge about sample origin.

However: SNPs make terrific markers for sample ID...

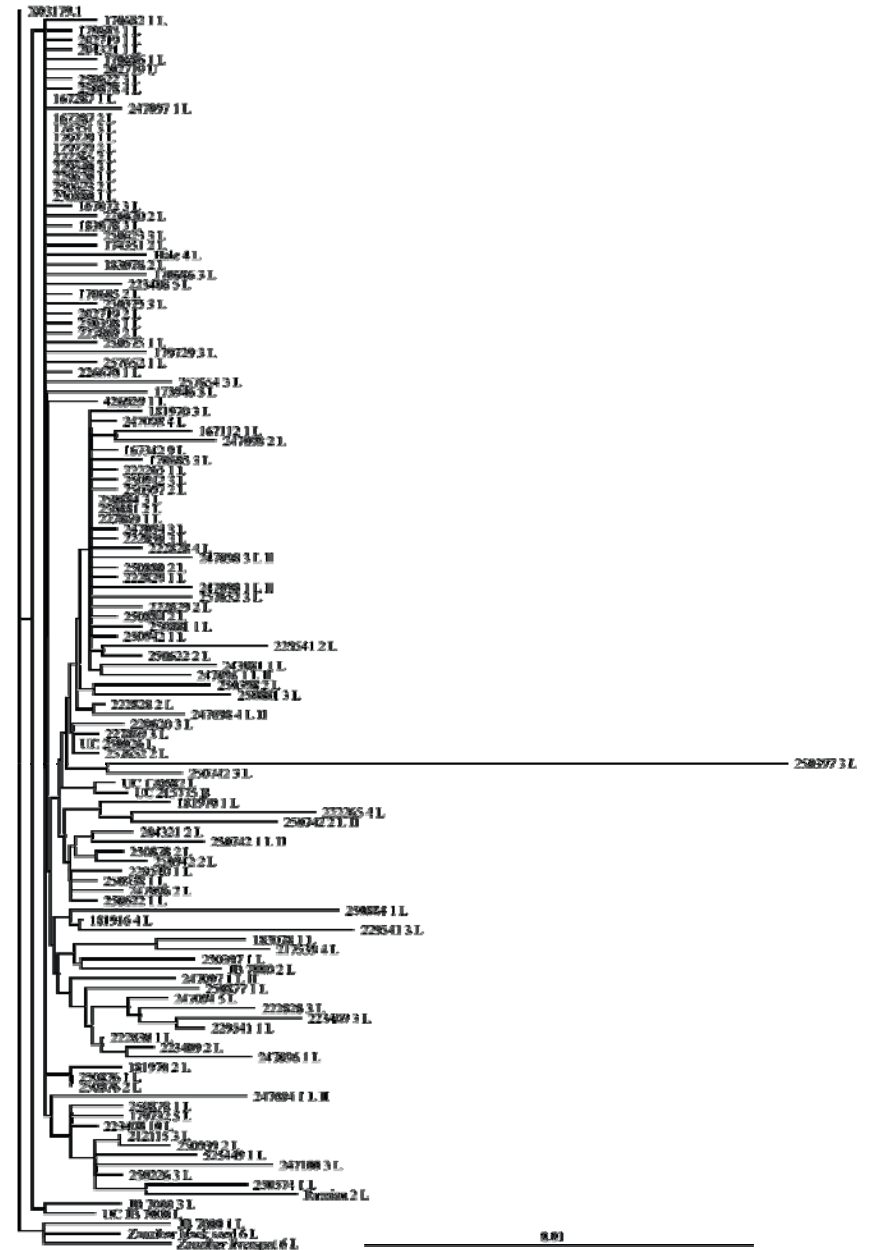
50% majority rule of 80,000
most-parsimonious trees

UNCLASSIFIED



UNCLASSIFIED

Analysis by neighbor joining
is also uninformative.



UNCLASSIFIED



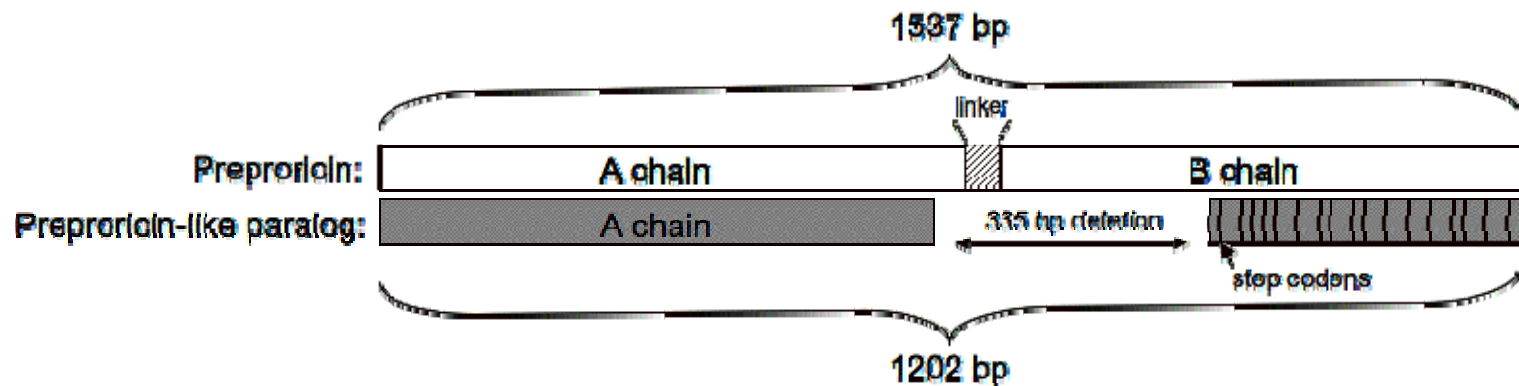
ECBC

Preproricin heterogeneity: small band

Sequenced 50 cloned “small” bands (1202 bp trimmed) from the 16 plants that yielded a small band.

ALL contained the same precise 335 base deletion relative to the “whole” preproricin sequence.

If this paralog is expressed, the A chain may retain function, but the deletion makes a frameshift that creates multiple missense and nonsense mutations.



Comparative nucleotide alignment of the small amplicon sequence (preproricin-like paralog) and the full-length preproricin sequence



UNCLASSIFIED

Preproricin heterogeneity: small band

Why do we think this is *not* an artifact of the PCR reaction?

The amplification is consistent

- same accessions yield/don't yield extra band

The deletion is precise

- non-specific priming would lead to a variety of sequences

- our primers bind appropriate targets

The occurrence among multiple accessions not cross-contamination

- SNPs occur among the paralogs, just like full length seqs

- if cross-contaminants, all sequences would be identical

The secondary structure of the deleted sequence

- modeling does not suggest an obvious structure

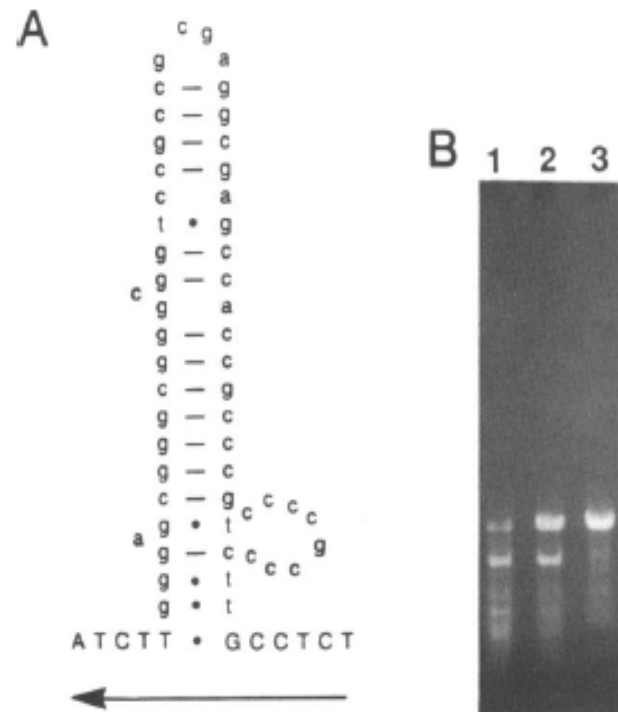


ECBC

UNCLASSIFIED



ECBC

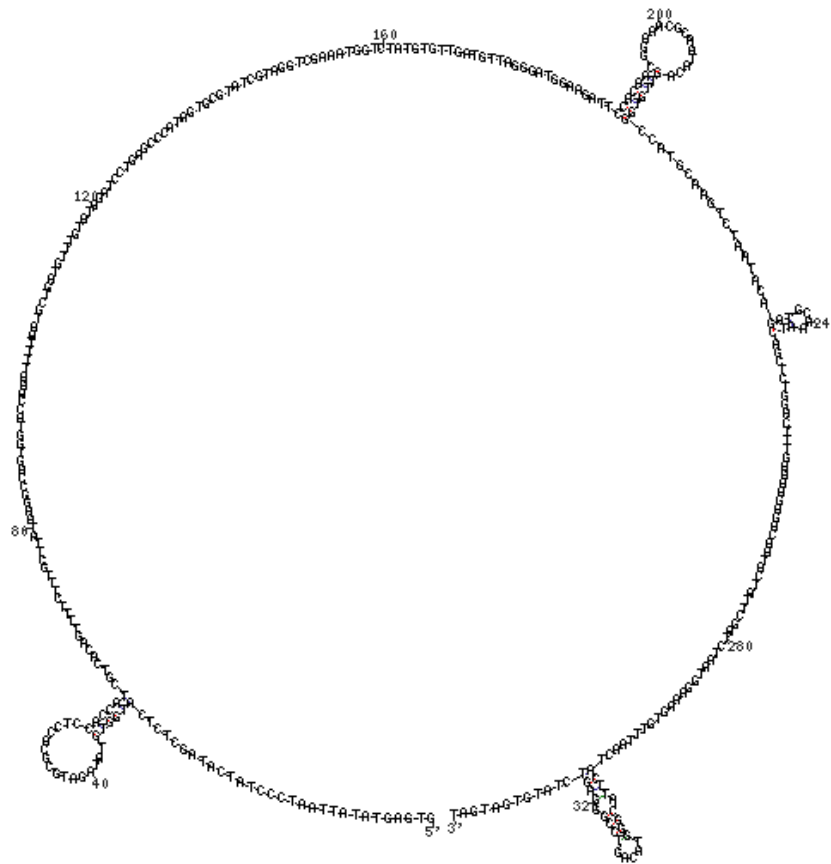


Mechanism for generating PCR-induced deletions (shown here in Human 18S rRNA)

(from Chou, *NAR* 20:4371, 1992)

UNCLASSIFIED

UNCLASSIFIED



dG = -4.50 06Nov20-16-07-49

Best structure of 335-bp deleted region modeled by mfold:
T=55C

No matter what ionic conditions entered, the ends of the deleted sequence cannot be made to basepair.

Forcing basepairing of an end generates structures with positive delta-G and still does not pair the other end precisely.

UNCLASSIFIED



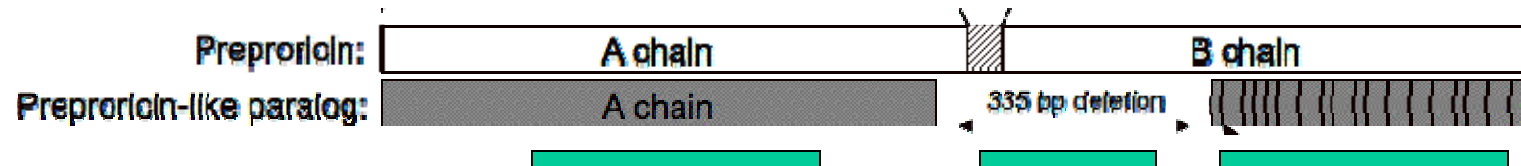
ECBC

Preproricin heterogeneity: small band

What's left? Two experiments:

- 1) PCR technique that discourages formation of small artifact bands
ss-DNA binding protein (NAR paper)
alternative buffer (Phusion GC buffer)
- 2) Southern hybridization with multiple probes

Are there bands that hybridize with probes from the 5' and 3' ends that do not with probes made from the 335-bp deleted region?



Summary:

- Sequenced 126 ~full-length genes from 63 accessions.
- Most sequences are unique (unshared SNPs)
- Not phylogenetically informative, but:
- Significant implications for sequence-based assays
- Obvious forensic applications
- The small paralog is previously undescribed
- Unlikely to be an artifact
- Defining an evolutionary event for RIP in *R. communis*
- Spread by sharing among breeding programs?
- Possible implications for anti-cancer therapeutics?



UNCLASSIFIED

To Do:

- Resequencing done; data being analyzed
- Research the source material (USDA records??)
- Southern blots: copy number (whole and small)
- AFLP analysis: assess genome-wide diversity
- Expand analysis to greater number of samples



The PI in
Tbilisi, Georgia

UNCLASSIFIED



ECBC

UNCLASSIFIED

Colleagues:

ECBC:

Andrea Weeks, Ph.D. (now @ George Mason U.)

Evan Skowronski, Ph.D.

Brian Kimble

Mike Horsmon

Georgetown University:

Jonathan Leshin

Ken Dretchen, Ph.D.



UNCLASSIFIED