Specification for LRGs: Locus Reference Genomic Sequences

Background

A meeting sponsored by GEN2PHEN (http://www.gen2phen.org) was held at the European Bioinformatics Institute (EBI), Hinxton, UK on 24–25 April 2008 to discuss a specification for stable reference genomic DNA sequences suited to the task of reporting variations. Attendees at the EBI meeting and their affiliations are listed in Appendix 1. The discussions took place within the context that NCBI, in collaboration with stakeholders, has already begun to provide gene-specific genomic sequences (RefSeqGene) that exactly complement its RefSeq mRNA and protein sequences.

In advance of the meeting, a survey was conducted by GEN2PHEN with the help of HGVS to assess the views of the curators of Locus-Specific Databases (LSDBs). The results of this survey are summarized in Appendix 2.

The primary goal of the meeting was to create a universally acceptable standard: a new specification for human genomic DNA Reference Sequences that addresses the primary shortcomings of existing systems, *viz.*:

- The lack of universally agreed genomic Reference Sequences for certain genes
- Inconsistent and incomplete (and sometimes outdated and inappropriate) annotation of the existing 'ad hoc' reference sequences
- DNA sequence inconsistencies between existing genomic reference sequences and their NCBI RefSeq mRNA sequence counterparts
- The lack of long-term stability of reference sequences
- Confusion in end-users' minds concerning 'versioning' of DNA sequences

Principles

The primary principles guiding the discussions with respect to the specifications for genomic reference sequences were:

- The sequences need not represent real alleles of genes: they can be composites that provide a practical working framework for the reporting of variations
- The community (*e.g.* research or diagnostic labs, LSDB curators, variation consortia, *etc.*) will have the final say in defining the sequences and their annotation
- Stability of the sequences, their core annotation, and their identifiers is paramount to ensure consistency of variation reporting over time frames of many decades

The Solution

The agreed solution was the concept of an LRG (Locus Reference Genomic), which builds on the initial ideas from NCBI for RefSeqGene. Specifically:

- LRG is a system for providing a genomic DNA sequence representation of a single gene that is idealised, has a permanent ID (with no versioning), and core content that never changes (*i.e.* nucleotide sequence, transcripts, exons, start & stop codon positions)
- LRG is not a nomenclature system (that is provided by the HGVS variation nomenclature)
- More than one LRG could be created for a region of interest, should that need arise

- Additional annotations will be present that may change with time (each item carrying its own date stamp), so that the latest ancillary knowledge about a gene is directly available. In other words, an LRG sequence and its core annotation are not meant to represent current biology knowledge, but to provide a standard for reporting variation in a stable coordinate system. The combination of the LRG plus the updatable-annotation layer will be used to support the biological interpretation of variants

The Details

A. Sequence Coordinates and Annotation

An LRG will be assigned a permanently stable identifier for the combination of a genomic sequence plus its core 'locked' annotation.

1. What defines the sequence content and coordinates?

- the sequence presented in transcriptional orientation
- sufficient 5' and 3' flanking sequence for unique placement in the genome
- by default, flanking sequences will be 5 kb upstream and 2 kb downstream of the identified transcript start and end points, although this can be made shorter or longer based on the wishes of the community requesting the sequence definition
- only one DNA strand will be represented in an LRG, the first base of which is numbered '1'
- no nucleotide ambiguity codes are allowed in the DNA sequence
- overlapping LRGs may be generated for any particular genome region, if good reason for this exists (*e.g.*, to support overlapping genes, or to accommodate newly identified nearby exons)
- after an LRG is produced for a gene, if a new exon for that gene happens to be identified that resides a long way outside the bounds of that LRG, it is proposed that a new LRG should be produced that spans just the extent of the new exon plus sufficient flanking sequence to place it in the genome

2. What is in the fixed-annotation layer in an LRG?

LRGs will be written in the DNA alphabet and will encompass a single gene. Overlapping genes encoded on the other strand, and hence transcribed in the opposite direction, will require their own separate LRGs.

A gene might give rise to one or more transcripts. In the context of an LRG, the term "transcript" means a fully processed functional RNA that is either coding (*i.e.*, an mRNA) or is non-coding (*e.g.*, tRNA or long & short ncRNAs). To avoid any confusion, "transcript" is not synonymous with a primary transcription product (*e.g.*, hnRNA). Different transcripts from a gene might share exons, or regions of exons, in common. Stakeholders will decide which transcripts are annotated in the fixed-annotation section of the LRG.

For each transcript "t" (numbered, sequentially, with an Arabic numeral (*e.g.*, t1, t2, t3, *etc.*)) the following information will be annotated:

- The sequence coordinates comprising the transcript
- For coding transcripts, amino acids will be numbered sequentially from the start codon with:
    - coordinates for the start codon

    - indication of selenocysteine (U) and pyrrolysine (O) codons
    - coordinates for the stop codon
- The conceptual translation protein sequence
- Non-coding transcripts (ncRNAs) will be annotated consistently with INSDC standards
- Species /tax_id (the LRG concept is currently motivated by the needs of human geneticists, but other species may subsequently be supported)
- Translation table (if different from standard translation table)
- Creation date
- Molecule type
- Plus any other standard attribute used by INSDC to define a sequence
- URL for LRG home page (i.e. http://www.lrg-sequence.org

In the fixed layer, the gene will be defined by placement of standard transcripts. The exons so placed will not be assigned explicit identifiers (labels or names), but will be defined by their coordinates on the LRG, the cDNA, and the coding region. The updatable layer, however, can be used to represent additional information about the same exons, or additional exons, thus allowing for both systematic and legacy numbering according to the needs of the stakeholders.

In summary, the fixed-annotation layer contains just the DNA sequence, the coordinates and sequence of the major transcripts, the coordinates of the exons that comprise these transcripts and the sequence of each conceptual translated protein. No other information concerning the biology of the gene is recorded in the fixed-annotation layer.

3. What is in the updatable-annotation layer in an LRG?

This layer, which will be updated when necessary, provides coordinates to map the LRG onto the current genome build, onto legacy reference sequences and to support legacy exon- & amino-acid-numbering systems.

The sequence of an LRG is intended to be stable and not change whenever human genome assemblies are revised and updated. In addition, the sequence of an LRG may be a composite of several natural alleles. Hence, it will be necessary to provide:

- coordinates and other information to map an LRG onto the human genome to allow for changes to reference genome builds:
    - for different assemblies (reference, HuRef, Celera)
    - history of versions of these assemblies (NCBI35/HG17|NCBI36/HG18)
    - these alignments and placements to be agreed on by NCBI/EBI/UCSC
- exon numbering scheme: Exons will be numbered according to the needs of the stakeholders by reference to the exon coordinates in the fixed-annotation layer. Ideally, the numbering will be sequential with Arabic numerals beginning at 1. Individual exons may be divided into sub-regions (e.g., 2a, 2b, 2c, etc.) to allow for complex RNA splicing events where splice donor or acceptor sites for certain transcripts lie wholly within complete exons of other transcripts (see Appendix 3 for a simple theoretical example)
- legacy DNA reference sequences and legacy exon- & amino-acid-numbering systems: It is important to recognise that variations may have already been described in the literature using legacy DNA reference sequences and legacy exon- & amino-acid-numbering systems. Coordinates will be included to map such legacy systems onto an LRG. This will

allow gene viewers to be developed that will ease the inter-conversion of variation data between legacy and LRG nomenclatures and coordinates

In addition, the updatable-annotation layer will contain:

- Audit data for each feature (creation date/modification date/data source)
- Chromosome number and location
- Additional features to be included
    - HGNC ID
    - HGNC gene symbol
    - ENSEMBL Gene
    - Entrez GeneID
    - OMIM geneID
    - regulatory regions
    - cross reference to RefSeqGene NG
    - cross references to RefSeq & ENSEMBL NM, NP, NR, ENST, ENSP by t{n} label
    - cross reference to any other legacy reference sequences
    - stakeholder contribution (attribution to LSDBs and contacts at LSDBs)
    - coordinates of additional genes, or parts of genes, that overlap, lie adjacent to, or are encoded on the opposite strand, relative to the primary gene of the LRG
    - inter-LRG tracking when regions overlap or LRGs are added
    - citations explicitly related to the LRG development

As new biological information becomes available, it may be necessary to annotate additional functional transcripts for the purposes of variation description. In the first instance, such transcripts will be added to the updatable-annotation layer. If these transcripts subsequently prove to be essential for variation description, they might be promoted to fixed-annotation layer.

4. Scope

- LRGs are nuclear only, because MitoMap manages the mitochondrial genome
- LRGs can be generated for isolated, proven regulatory regions (*i.e.* without a gene feature)

5. Scenarios demanding a new LRG: (rare, if a well-designed LRG already exists for a locus)

- Need to change genomic sequence
- Need to represent a different structural haplotype
- Error made in identifying any core ‘locked’ annotation

B. Implementation

There will be joint EBI/HGNC/NCBI implementation of LRGs to ensure that gene IDs and symbols are correctly assigned. Journal editors and LSDBs curators will be informed of LRGs with the intention of raising awareness of the advantages of using LRGs as reference sequences. It is hoped that journals might mandate the use of LRGs in the description of gene variants.

C. LRG numbering and reporting conventions

- LRGs will be numbered sequentially with Arabic numerals with no leading zeros to pad the number to a fixed length (*i.e.* LRG_1, LRG_2, LRG_3, *etc.*)
- There will be no versioning of LRGs
- For transcripts, even if only 1 is known, indicate this by LRG_{n}t1
- If no transcripts are known, then LRG_{n}t1 is not required
- Additional transcripts named by sequential integers (*i.e.* LRG_{n}t2, LRG_{n}t3, *etc.*)
- Reporting conventions, using HGVS nomenclature:
    - LRG_1:g.4G>C (genomic position numbering)
    - LRG_1t1:c.24C>T (cDNA position numbering, relative to first nucleotide of start codon)
    - LRG_1t2:63T>A (non-coding transcript position numbering, starts at first nucleotide of transcript: only allowed when no CDS present)
    - People may also want to report the protein variant. The recommendation is that this be reported along with a nucleotide report,
      e.g. LRG_1t1:c.572G>A (p.Gly191Asp)

D. Outline of LRG production process

A. Identify data stakeholders (hope many are self-identifying), and with them generate an initial LRG proposal. If disagreement results in trying to achieve the initial definition, one will be selected by the LRG production committee

B. Provide all stakeholders with information about the proposed LRG and request review of proposed sequence and associated annotations

C. Once everyone is happy with the proposal, identify the next available LRG number and assign

D. Generate LRG formats for download:

- LRG: LRG DNA sequence and core (stable) annotations
- LRGplus: LRG DNA and protein sequences, and core and updateable annotations
    - all old releases maintained in archive
- Formats:
    - XML based on the INSDSeq XML standard by INSDC
    - other useful formats, *e.g.* genbank, embl, fasta, gff, *etc.* might be generated, if necessary, by transformation using XSL style sheets

Tools

- Tools will be required to visualise LRGs and their relationship to previous reference sequences (*e.g.* RefSeq and RefSeqGene entries) and to other related LRGs, otherwise the community acceptance of the LRG standard might be limited. Decisions about tool features ought to be informed by community requests and the primary goal should be ease of use. Such tools could be based on NCBI Genome Workbench, Ensembl or the NGRL variant browser, but a tool dedicated to LRGs and variation reporting would be desirable. Functionality should include:
    - Select the appropriate LRG by HGNC gene symbol

    - Display of alignment of previous reference to current
    - Alignment of DNA and protein sequences
    - Ability to edit sequence in a “what if” fashion and view the consequences in terms of alterations to translation and/or splicing using HGVS-compliant nomenclature
- Mutalyzer, Mutation Checker and other similar tools will need to be adapted to parse LRGs

Glossary of Abbreviations and Terms

The following abbreviations and terms have been used:

- **EBI**: European Bioinformatics Institute (http://www.ebi.ac.uk/)
- **ENSEMBL**: The EBI/Sanger Institute genome browser (http://www.ensembl.org/)
- **GEN2PHEN**: Genotype-To-Phenotype Databases: A Holistic Solution; European Community Seventh Framework Programme, HEALTH Theme — Contract No. 200754 (http://www.gen2phen.org/)
- **HGNC**: HUGO Gene Nomenclature Committee (http://www.genenames.org/)
- **HGVS**: Human Genome Variation Society (http://www.hgvs.org/)
- **HUGO**: Human Genome Organisation (http://www.hugo-international.org/)
- **INSDC**: International Nucleotide Sequence Database Collaboration (http://www.insdc.org/)
- **INSDSeq**: An XML-based sequence format (http://www.ebi.ac.uk/embl/dtd/INSD_V1.4.dtd)
- **LRG**: Locus Reference Genomic
- **NCBI Genome Workbench**: an integrated application for viewing and analysing sequence data (http://www.ncbi.nlm.nih.gov/projects/gbench/)
- **NCBI**: National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/)
- **NGRL**: National Genetics Reference Laboratory, Manchester (http://www.ngrl.org.uk/Manchester/)
- **Mutalyzer**: a tool for checking sequence variant nomenclature (http://www.lovd.nl/mutalyzer/)
- **Mutation Checker**: a tool to verify the effects of DNA-level sequence variation (http://mutation.sanbi.ac.za/checker)
- **Pyrrolysine:** the 22nd amino acid (http://www.ncbi.nlm.nih.gov/pubmed/12029131)
- **RefSeq**: NCBI Reference Sequence (http://www.ncbi.nlm.nih.gov/RefSeq/)
- **RefSeqGene**: NCBI Reference Genomic Sequence (http://www.ncbi.nlm.nih.gov/RefSeq/RSG/)
- **Selenocysteine:** the 21st amino acid (http://www.ncbi.nlm.nih.gov/pubmed/12524431)
- **UCSC:** UCSC Genome Bioinformatics (http://genome.ucsc.edu/)

Appendix 1: Meeting Attendees

University of Leicester, Leicester

Raymond Dalgleish

Anthony J Brookes

EBI, Hinxton

Ewan Birney

Paul Flicek

Glenn Proctor

Elspeth Bruford

NCBI, Bethesda

Donna Maglott

Steve Sherry

Mike Feolo

INSERM, Montpellier

Marine Lalande

François-Olivier Desmet

Leiden University Medical Centre, Leiden

Peter Taschner

South African National Bioinformatics Institute, Cape Town

Heikki Lehväslaiho

BIOBASE, Wolfenbüttel

Tatiana Konovalova

National Genetics Reference Laboratory, Manchester

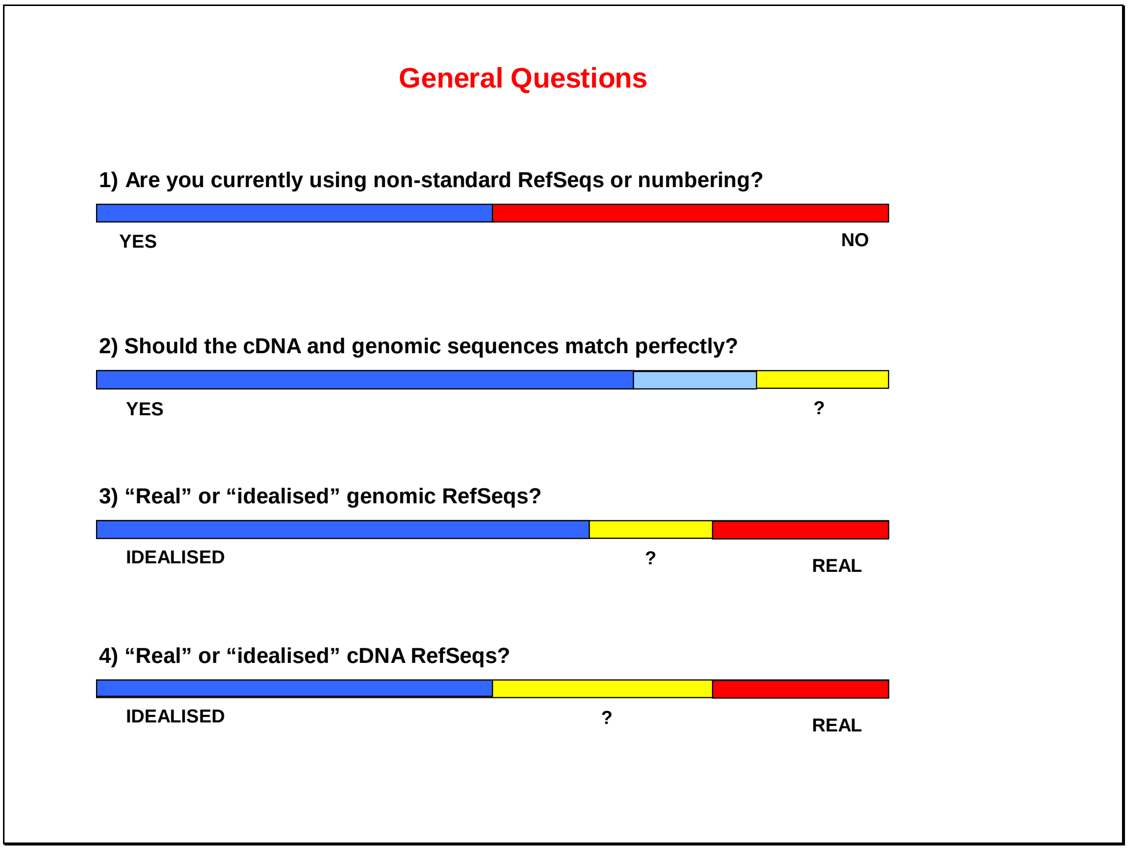
Ed Burke

In attendance

Myles Axton, Editor of Nature Genetics

Appendix 2: Summary of Survey Results

The following PowerPoint slides summarise the views of LSDB curators who responded to a survey conducted in March and April 2008. These results were used to inform the discussion that resulted in the proposal for LRGs.



General Questions

1) Are you currently using non-standard RefSeqs or numbering?

YES NO

2) Should the cDNA and genomic sequences match perfectly?

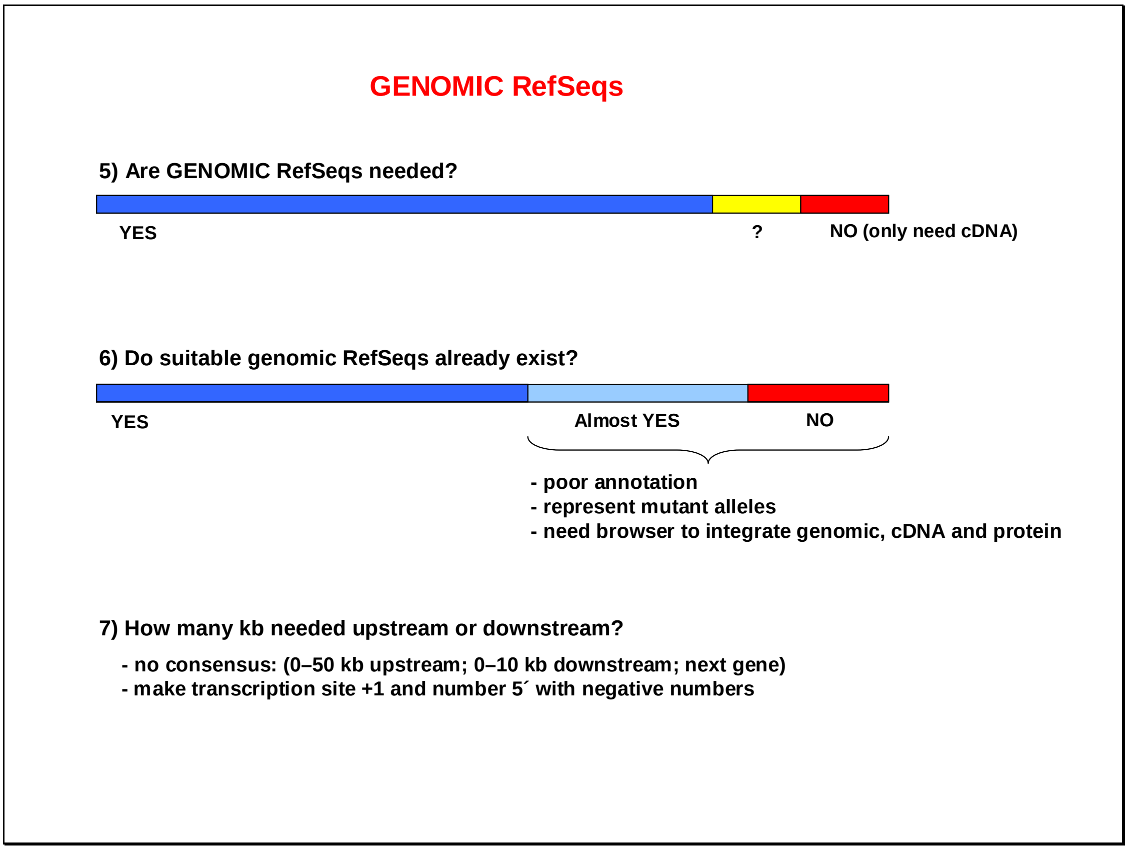
YES ?

3) “Real” or “idealised” genomic RefSeqs?

IDEALISED ? REAL

4) “Real” or “idealised” cDNA RefSeqs?

IDEALISED ? REAL



GENOMIC RefSeqs

5) Are GENOMIC RefSeqs needed?

YES ? NO (only need cDNA)

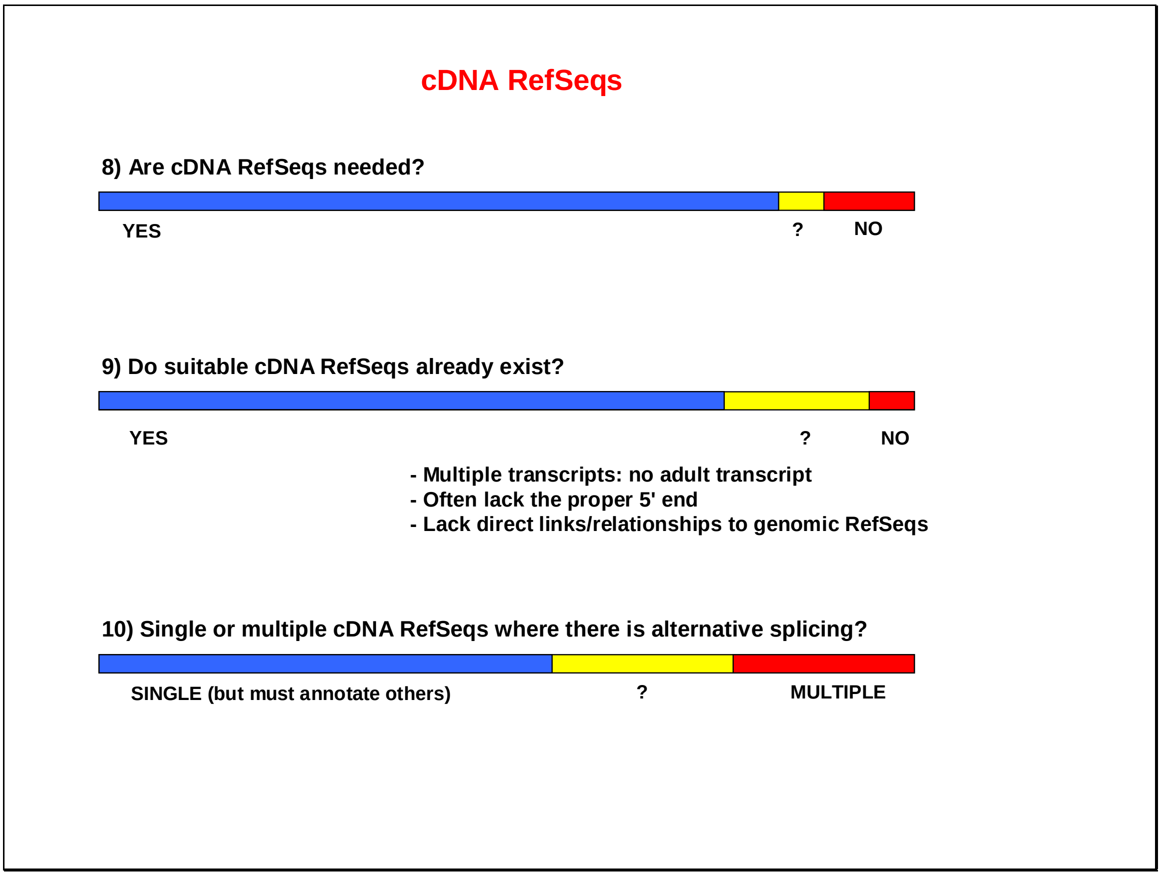
6) Do suitable genomic RefSeqs already exist?

YES Almost YES NO

- poor annotation
- represent mutant alleles
- need browser to integrate genomic, cDNA and protein

7) How many kb needed upstream or downstream?

- no consensus: (0–50 kb upstream; 0–10 kb downstream; next gene)
- make transcription site +1 and number 5´ with negative numbers



cDNA RefSeqs

8) Are cDNA RefSeqs needed?

YES ? NO

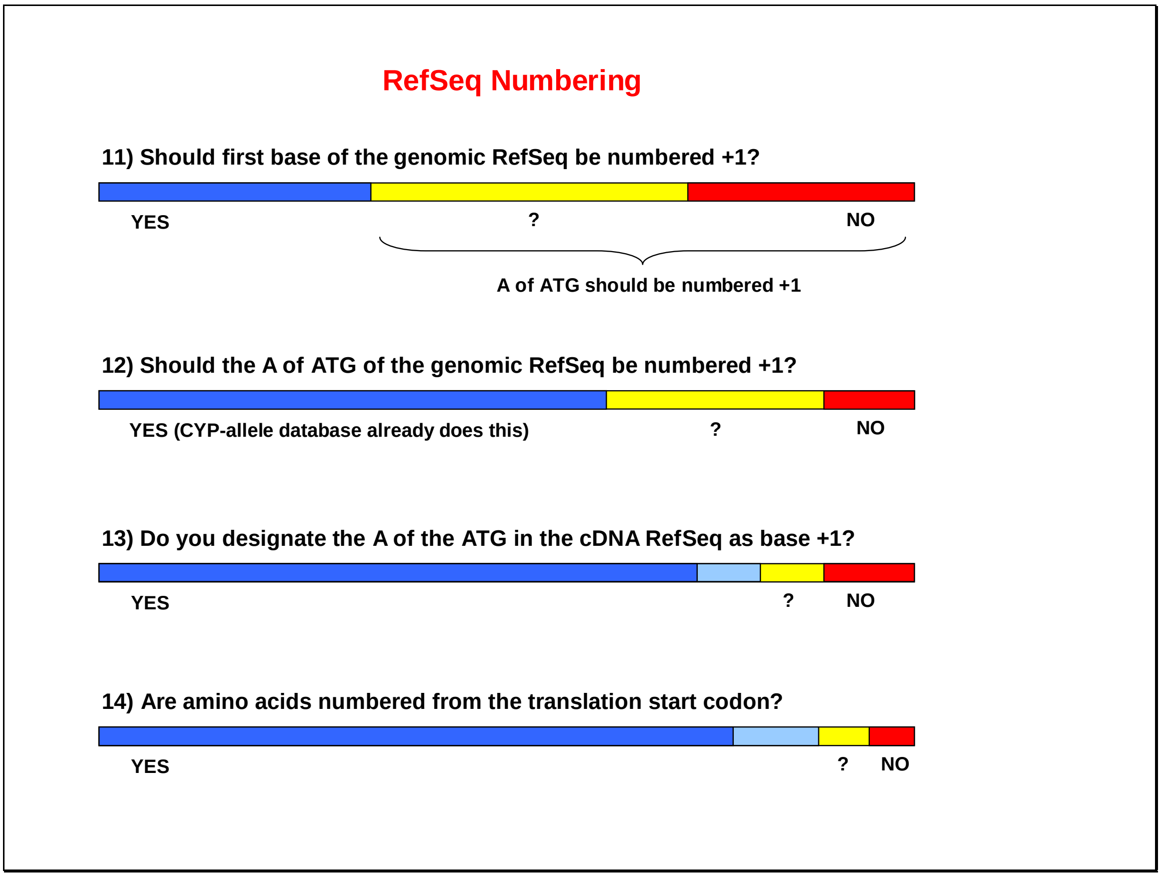
9) Do suitable cDNA RefSeqs already exist?

YES ? NO

- Multiple transcripts: no adult transcript
- Often lack the proper 5' end
- Lack direct links/relationships to genomic RefSeqs

10) Single or multiple cDNA RefSeqs where there is alternative splicing?

SINGLE (but must annotate others) ? MULTIPLE



RefSeq Numbering

11) Should first base of the genomic RefSeq be numbered +1?

YES ? NO

A of ATG should be numbered +1

12) Should the A of ATG of the genomic RefSeq be numbered +1?

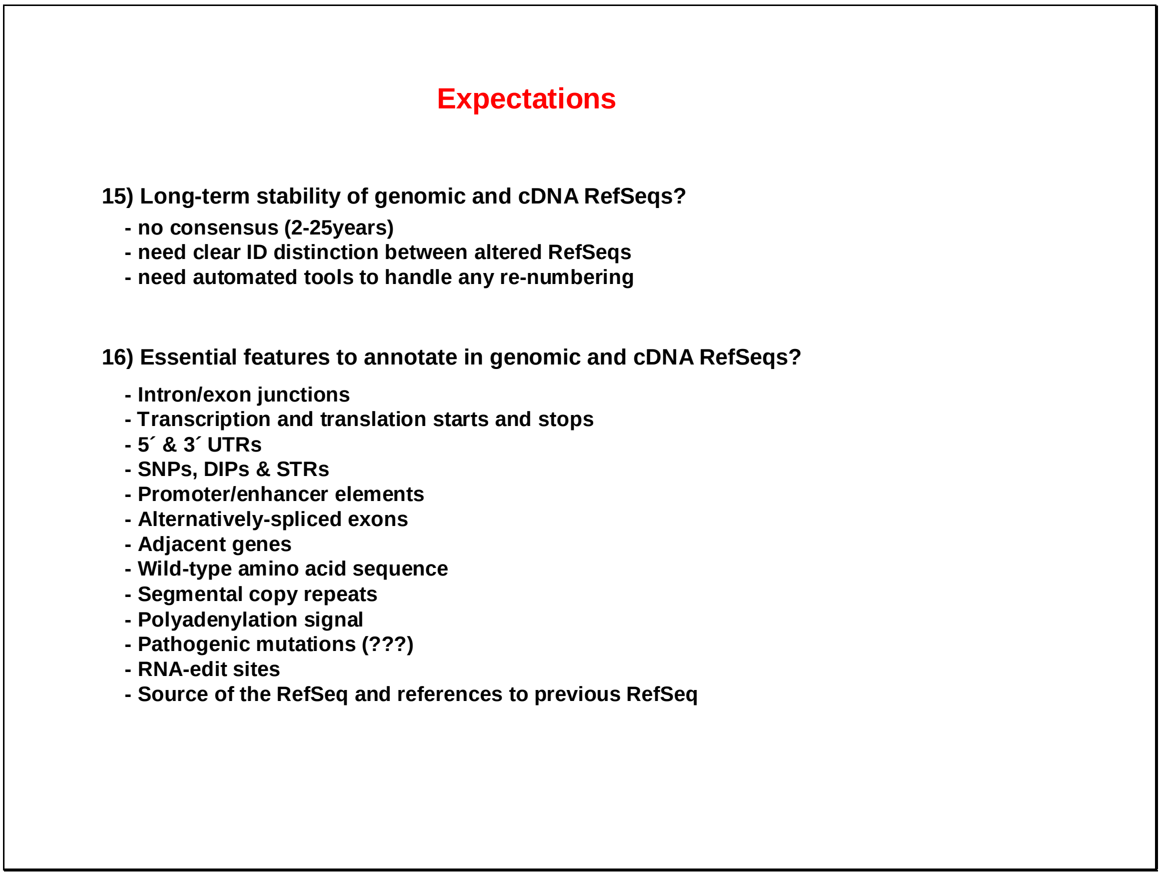
YES (CYP-allele database already does this) ? NO

13) Do you designate the A of the ATG in the cDNA RefSeq as base +1?

YES ? NO

14) Are amino acids numbered from the translation start codon?

YES ? NO



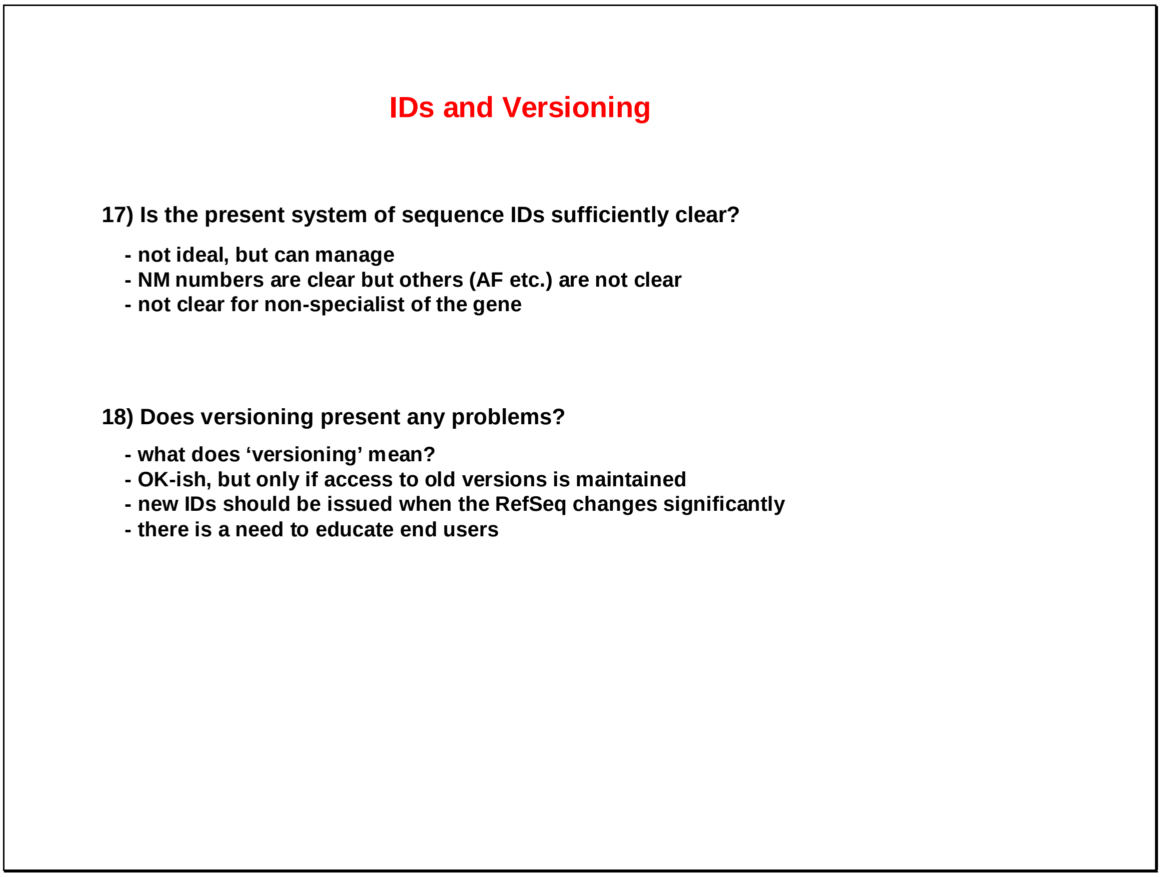
Expectations

15) Long-term stability of genomic and cDNA RefSeqs?

- no consensus (2-25years)
- need clear ID distinction between altered RefSeqs
- need automated tools to handle any re-numbering

16) Essential features to annotate in genomic and cDNA RefSeqs?

- Intron/exon junctions
- Transcription and translation starts and stops
- 5´ & 3´ UTRs
- SNPs, DIPs & STRs
- Promoter/enhancer elements
- Alternatively-spliced exons
- Adjacent genes
- Wild-type amino acid sequence
- Segmental copy repeats
- Polyadenylation signal
- Pathogenic mutations (???)
- RNA-edit sites
- Source of the RefSeq and references to previous RefSeq



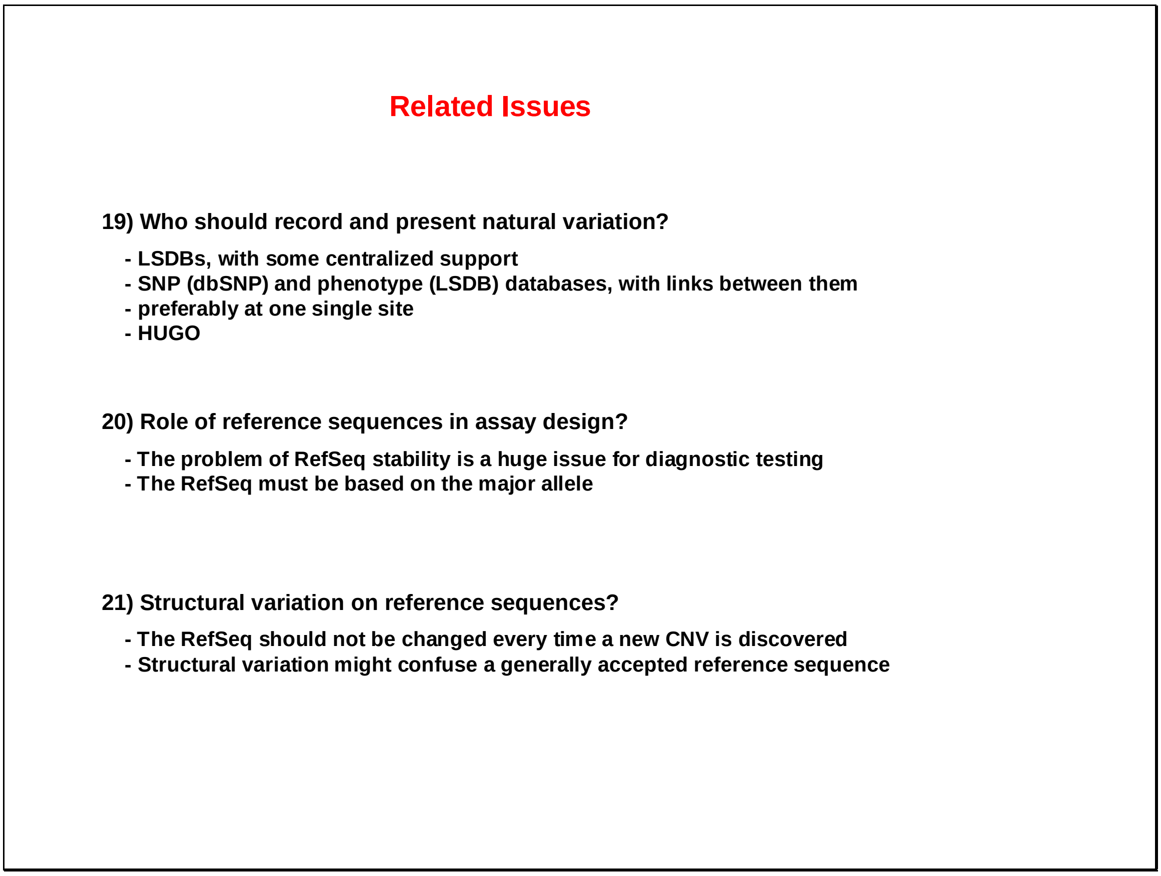
IDs and Versioning

17) Is the present system of sequence IDs sufficiently clear?

- not ideal, but can manage
- NM numbers are clear but others (AF etc.) are not clear
- not clear for non-specialist of the gene

18) Does versioning present any problems?

- what does ‘versioning’ mean?
- OK-ish, but only if access to old versions is maintained
- new IDs should be issued when the RefSeq changes significantly
- there is a need to educate end users



Related Issues

19) Who should record and present natural variation?

- LSDBs, with some centralized support
- SNP (dbSNP) and phenotype (LSDB) databases, with links between them
- preferably at one single site
- HUGO

20) Role of reference sequences in assay design?

- The problem of RefSeq stability is a huge issue for diagnostic testing
- The RefSeq must be based on the major allele

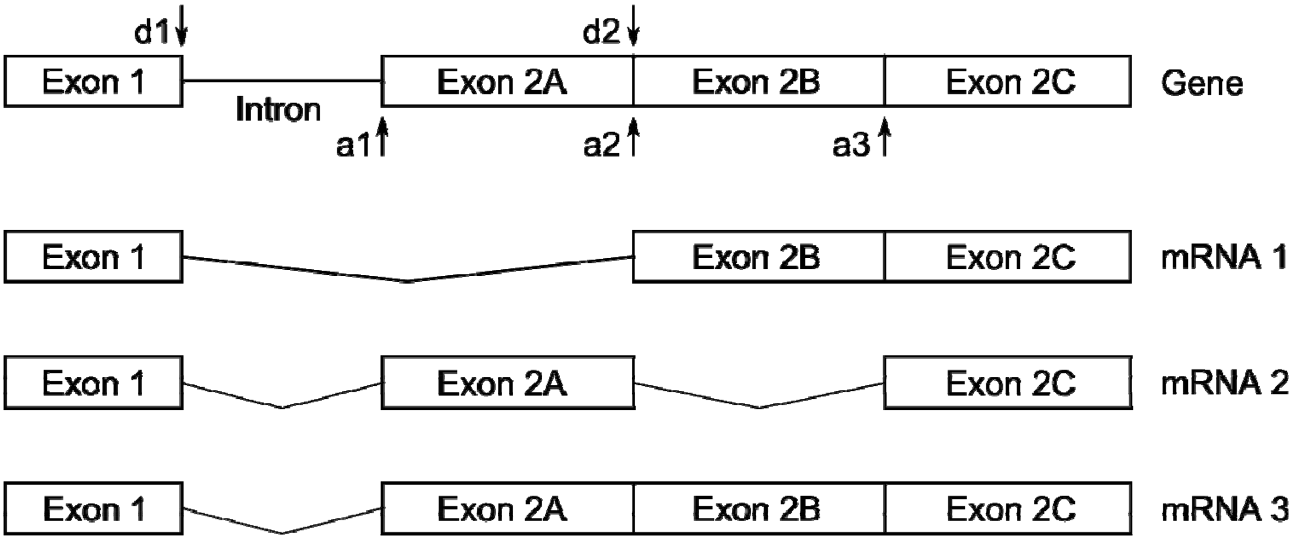
21) Structural variation on reference sequences?

- The RefSeq should not be changed every time a new CNV is discovered
- Structural variation might confuse a generally accepted reference sequence

Appendix 3: Exons with internal splice sites

The diagram below represents exons 1 and 2 of a theoretical gene. Exon 2 may be spliced directly to exon 1, yielding mRNA 3. Exon 2 contains an internal splice donor site (d2) and two splice acceptor sites (a2 & a3) allowing for mRNAs 1 & 2 to be produced by alternative splicing.

Exons 2A, 2B and 2C are deemed to comprise exon 2 because their sequences are contiguous.


d1
d2
Exon 1
Intron
Exon 2A
Exon 2B
Exon 2C
Gene
a1
a2
a3
Exon 1
Exon 2B
Exon 2C
mRNA 1
Exon 1
Exon 2A
Exon 2C
mRNA 2
Exon 1
Exon 2A
Exon 2B
Exon 2C
mRNA 3