developing countries are actively developing policy to engage with GM crops, there is indeed very little going on in terms of GM insects, which, for the record, will ignore national boundaries. An international entity with broad, adaptive and adequate representation is therefore urgently called for. Given the right mandate, it can safeguard against uncontrolled expansion of activities while serving as a shield for antagonistic influences through active stakeholder engagement.

Finally, following the foregoing multiple perspective debates on GM mosquitoes, we propose the rapid initiation of an international gathering to start addressing the complexity of ethical, legal and social aspects of GM mosquitoes for disease control, a process that should already have taken place[16,17]. We conclude that contrary to there being a 'green light for mosquito control,' as announced in your journal[18], research on SIT using transgenic insects has, for now at least, stalled at a yellow light.

Bart G J Knols[1], Rebecca C Hood-Nowotny[1], Hervé Bossin[1], Gerald Franz[1], Alan Robinson[1], Wolfgang R Mukabana[2] & Samuel K Kemboi[2]

[1]Entomology Unit, FAO/IAEA Agriculture and Biotechnology Laboratory, A-2444 Seibersdorf, Seibersdorf, Vienna, Austria. [2]University of Nairobi, P.O. Box 29053, Nairobi, Kenya. e-mail: B.Knols@iaea.org

1. Dyck, A.V., Hendrichs, J. & Robinson, A.S. (eds.) The Sterile Insect Technique: Principles and Practice in Area-Wide Integrated Pest Management (Springer, Heidelberg, 2005).
2. Catteruccia, F. *et al. Science* **299**, 1225–1227 (2003).
3. Andreasen, M. & Curtis, C.F. *Med. Vet. Entomol.* **19**, 238–244 (2005).
4. Franz, G. *Genetica* **116**, 73–84 (2002).
5. Benedict, M. & Robinson, A.S. *Trends Parasitol.* **19**, 349–355 (2003).
6. Scott, T.A., Takken, W., Knols, B.G.J. & Boete, C. *Science* **298**, 117–119.
7. Alphey, L. *et al. Science* **298**, 119–121 (2002).
8. Takken, W. & Scott, T.A. (eds.) *Ecological Aspects for Application of Genetically Modified Mosquitoes.* (Kluwer Academic Publishers, Dordrecht, The Netherlands, 2005) <http://library.wur.nl/frontis/malaria/>
9. Knols, B.G.J. & Louis, C. (eds.) *Bridging Laboratory and Field Research for Genetic Control of Disease Vectors* (Springer, Berlin, 2005). <http://library.wur.nl/frontis/disease_vectors/>
10. The Royal Society. Risk Analysis, Perception and Management. Report of the Royal Society Study Group (The Royal Society, London, 1992).
11. Wynn, B. *Global Environ. Change* **June,** 111–127 (1992).
12. Rondinelli, D. *Development Projects as Policy Experiments.* (Routledge, London & New York, 1993).
13. Lusk, J.L. & Rozan, A. *Trends Biotechnol.* **23**, 386–387 (2005).
14. World Health Organisation. *WHO Chronicle* **30**, 131–139 (1976).
15. Ison, R.L. *Rangeland J.* **15**, 154–166 (1993).
16. Macer, D. Ethical, Legal and Social Issues of Genetically Modified Disease Vectors in Public Health. TDR/STR/SEB/ST/03.1 (World Health Organisation, Geneva, Switzerland, 2003).
17. Touré, Y.T. & Knols, B.G.J. in *Genetically Modified Mosquitoes for Malaria Control* (Boëte, C., ed.) (Landes Bioscience, Georgetown, Texas, USA, in the press, 2006).
18. Atkinson, P. *Nat. Biotechnol.* **23**, 1371–1372 (2005).

Peter Atkinson responds:

Knols *et al.* draw attention to two important points: that any new genetic strain developed for use in the sterile insect technique must undergo rigorous testing to ensure that it meets the necessary quality control standards required for the successful application of this technique; and that there must be full consultation with the public, stakeholders and any other interested parties before transgenic strains can be released. These self-evident facts are not in dispute; rather, the advance reported by Crisanti and colleagues in *Nature Biotechnology* illustrates that recombinant techniques are now generating genetic strains that may now be appropriate for assessment and, pending the outcome, deployment in insect genetic control programs. The application of these developments do need to be openly discussed in the type of forum outlined by Knols et al. and, toward this goal, preliminary workshops on this topic have already been convened[1].

1. Takken, W. & Scott, T.W. (eds.) *Ecological Aspects for Application of Genetically Modified Mosquitoes. Reports from a Workshop held at Wageningen University and Research Center, June 2002* (Kluwer Academic Publishers, Dordrecht, The Netherlands, 2003).

Sequencing errors or SNPs at splice-acceptor guanines in dbSNP?

To the editor:

Single-nucleotide polymorphisms (SNPs) are the most frequent type of human genetic variation. They are the major basis of our phenotypic individuality, particularly with respect to heritable differences in disease susceptibility. Large collections of mapped SNPs, public and private, are powerful tools for genetic studies[1]. The most comprehensive public SNP database, dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP), currently contains more than 12 million human SNPs (version 126). This wealth of data is extensively used by a broad community, including clinical, experimental and computational scientists, for both locus-specific and genome-wide studies. Therefore, the quality and completeness of dbSNP is of paramount importance and a recent meta-analysis of four confirmation studies estimated a false-positive rate of ~15–17%[2].

As we have an interest in alternative splicing in general[3] and with respect to diseases in particular[4], we searched dbSNP for human variations in a nine-nucleotide context (three exon and six intron positions) of all splice-donor/acceptor sites of mRNA RefSeqs. Contrary to our expectation for the highly conserved intron positions +1, +2 (donor) and –2, –1 (acceptor), the acceptor G at –1 showed a variability comparable to that of the random position –4 (**Fig. 1a**). As the disruption of the G at –1 normally results in the loss of the acceptor site[5], we questioned whether this surprising variability could be compensated by any of the known biological processes (for example, RNA editing) or is an indication for a yet unknown biological phenomenon. As we could not shape a plausible explanation for our observation, and before we considered undertaking a challenging, lengthy and potentially fruitless search for an unknown biological mechanism, we decided next to evaluate the possibility that false-positive entries in dbSNP are accountable for the inexplicable variability of position –1.

To this end, we first used the dbSNP validation status description and classified the RefSNPs (dbSNP entries) in three categories: (C1) validated by frequency or genotype data from HapMap[6] or any other submitter; (C2) validated by independent submissions, observation of the minor allele in at least two chromosomes or submitter confirmation; and (C3) single submission without confirmation. Conspicuously, position –1 showed the highest fraction in C3 (305 of 364, 84%; **Fig. 1b**). As experimental verification of RefSNPs depends on the availability of appropriate population samples and assays, it was not feasible for us to carry out such a study on a large scale. Therefore, we switched to a verification procedure making use of the electropherograms derived from automatic fluorescence-based DNA sequencing instruments (traces).

© 2006 Nature Publishing Group http://www.nature.com/naturebiotechnology

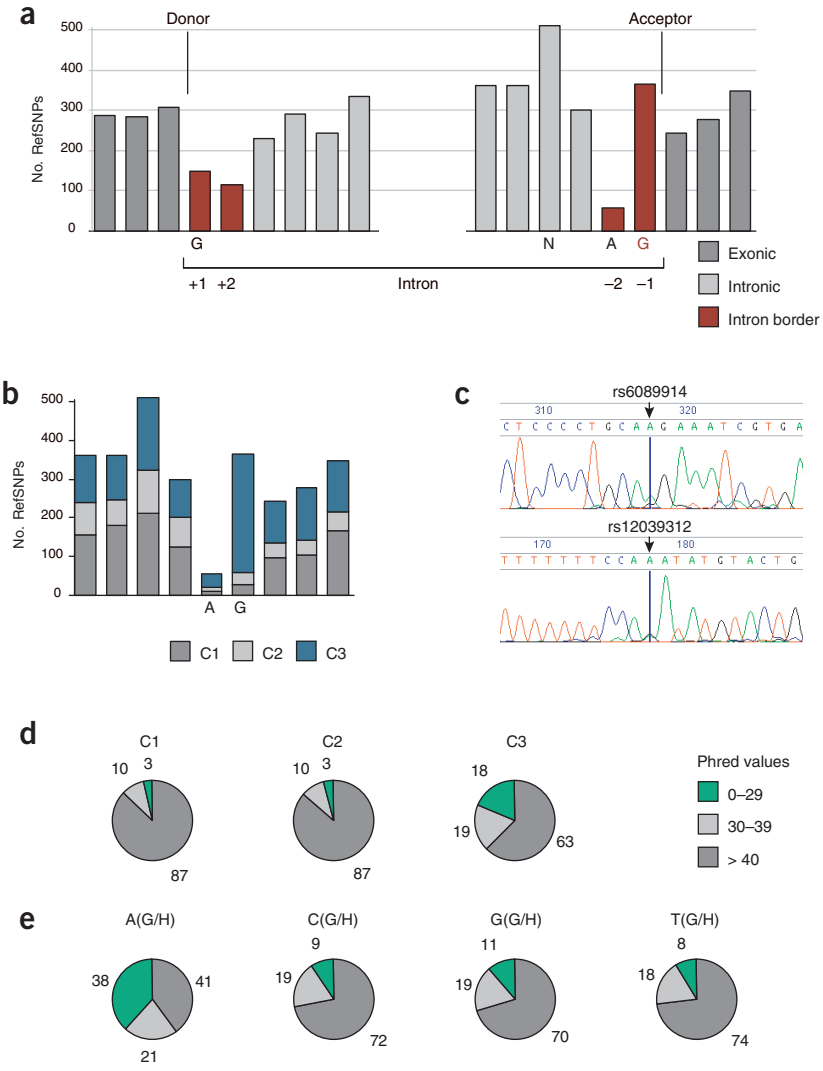

Figure 1 RefSNPs and sequence confidence. (**a**) Apparent hypervariability at splice-acceptor Gs. (**b**) Classification of RefSNPs at the splice acceptors according to their validation status. (**c**) Electropherograms (traces) illustrating the 'G after A' problem at splice-acceptor sites in the 5′-to-3′ sequencing direction. (**d**,**e**) Sequence confidence (Phred) values of trace data supported RefSNPs (**d**) classified according to their validation status and of G/H RefSNPs (**e**) classified according to the 5′ nucleotide; (**d**,**e**) numbers expressed as a percentage

Currently, 76% of all RefSNPs are supplied with trace references and for nearly 60% these data are accessible via the US National Center for Biotechnology Information (NCBI) Trace Archive (http://www.ncbi.nlm.nih.gov/Traces; **Supplementary Notes**). We manually examined the available traces for RefSNPs at acceptor positions –2, –1 and +1 and collected false-positive entries, which we classified as sequencing errors (wrong base calling due to low signal-to-noise ratio) and database errors (identity of genomic RefSeq and the trace supported RefSNP allele or ambiguous alignment in microsatellites). Sequencing errors were mainly detected among C3 RefSNPs that are solely based on single-pass trace data. Database errors occurred both in C2 and C3 RefSNPs independently of their trace coverage (single trace, multiple traces of the same strand, traces from both strands; **Supplementary Notes** online).

The astonishing error rate of 93% among 181 RefSNPs with trace data at acceptor position –1 was exclusively caused by the well-known suppression of G after A incorporation using thermostable, genetically engineered DNA polymerases in dye terminator sequencing reactions[7] (**Fig. 1c**). Naturally, this problem occurs at acceptor sites only in forward (5′-to-3′) traces because the AG is CT in the reverse sequencing direction. Moreover, the 'G after A' problem is further enhanced by the polypyrimidine tract preceding the acceptor AG in the splice consensus[8]. Homopolymer stretches of T and C are known to cause problems with sequence accuracy as a result of polymerase slippage[9], thus leading to elevated error rates not only at position –1 but also at –2 and +1.

Altogether, we estimated false-positive rates at acceptor positions –2, –1 and +1 of 17%, 82% and 11%, respectively (**Supplementary Tables 1–3** online). Excluding the estimated false-positive rates, no significant difference in the variability between acceptor positions –1 and –2 remains. Thus, we conclude that a systematic sequencing error ('suppressed G after A') and not a previously unknown biological phenomenon causes the high frequency of RefSNPs in splice-acceptor position –1.

Sensitized by this analysis, we then asked to what extent dbSNP contains sequencing errors in general. First, a scan of all RefSNPs for the sequence confidence of the allele alternative to the genomic RefSeq confirmed our initial observation that false positives are very likely enriched among C3 entries (18% with Phred confidence value <30; which means more than one error among 1,000 entries[10]) and will be equally rare among C1 and C2 entries (**Fig. 1d**; **Supplementary Notes** online). Moreover, the 'suppressed G after A' problem is not restricted to acceptor sites because among all G/H (genomic RefSeq allele/non-RefSeq allele, where H stands for A,C or T) C3 RefSNPs with traces, the fraction of low-confidence entries among A(G/H) variations is twice as large as for the remaining contexts (**Fig. 1e**; **Supplementary Notes** online).

For a concluding estimation of sequencing errors in dbSNP, we selected a set of 10,000 random SNPs and manually examined representative trace sets for all possible N(N/N)N contexts (where N is any nucleotide). Along with the expected A(G/H)N, the C(A/Y)C and G(A/C)C contexts also showed false-positive rates >10%. Altogether, we estimated that there were about 256,000 sequencing and 124,000 database errors, representing 3.2% and 1.5% of all RefSNPs. Among sequencing errors, the vast majority (85%) are caused by the 'suppressed G after A' problem. Most interestingly, some of the false RefSNPs were investigated in the HapMap project[6] (**Supplementary Tables 1–3** online) and, as expected, did not show any variation in all genotyped populations.

The described error rates in dbSNP might both introduce serious biases in large-scale bioinformatic studies and misdirect experimental efforts, particularly if a special sequence context such as acceptor AG is considered. Therefore,

© 2006 Nature Publishing Group http://www.nature.com/naturebiotechnology

we emphatically recommend all users of dbSNP to refer to the 'validation status' tag and use a simple SNP classification scheme, as described above, that aims at extracting RefSNPs with lower error rates. According to our classification, dbSNP (version 124) contains in C1, C2 and C3 2,077,680, 2,946,840 and 3,470,166 entries, respectively. To investigate the differences between those three classes, we extracted the available confidence information. C1 and C2 RefSNPs have higher average values (both 51.4) than SNPs in C3 (43.2, **Supplementary Notes** online). Furthermore, about 87% in C1 and C2 have confidence values of at least 40, in contrast to only 63% in C3 (**Fig. 1d**). As a low confidence value indicates a potential sequencing error, we recommend that bioinformatics and/or experimental efforts either use only C1 and C2 RefSNPs or find a way of excluding from C3 all dbSNP entries with Phred <40 (ref. 11).

Note: Supplementary information is available on the Nature Biotechnology website.

Matthias Platzer[1], Michael Hiller[2],
Karol Szafranski[1], Niels Jahn[1], Jochen Hampe[3], Stefan Schreiber[3], Rolf Backofen[2] & Klaus Huse[1]

[1]Genome Analysis, Leibniz Institute for Age Research–Fritz Lipmann Institute, Beutenbergstr. 11, 07745, Jena, Germany. [2]Institute of Computer Science, Albert-Ludwigs-University Freiburg, Georges-Koehler-Allee 106, 79110 Freiburg, Germany. [3]Institute for Clinical Molecular Biology, Christian-Albrechts-University Kiel, Schittenhelmstr. 12, 24105, Kiel, Germany. e-mail: mplatzer@fli-leibniz.de

1. Kruglyak, L. *Nat. Genet.* **17**, 21–24 (1997).
2. Mitchell, A.A., Zwick, M.E., Chakravarti, A. & Cutler, D.J. *Bioinformatics* **20**, 1022–1032 (2004).
3. Hiller, M. *et al. Nat. Genet.* **36**, 1255–1257 (2004).
4. Valentonyte, R. *et al. Nat. Genet.* **37**, 357–364 (2005).
5. Krawczak, M., Reiss, J. & Cooper, D.N. *Hum. Genet.* **90**, 41–54 (1992).
6. International HapMap Consortium. *Nature* **426**, 789–796 (2003).
7. Korch, C. & Drabkin, H. *Genome Res.* **9**, 588–595 (1999).
8. Stephens, R.M. & Schneider, T.D. *J. Mol. Biol.* **228**, 1124–1136 (1992).
9. Kotlyar, A.B., Borovok, N., Molotsky, T., Fadeev, L. & Gozin, M. *Nucleic Acids Res.* **33**, 525–535 (2005).
10. Ewing, B. & Green, P. *Genome Res.* **8**, 186–194 (1998).
11. Hiller, M. *et al. Am. J. Hum. Genet.* **78**, 291–302 (2006).

Data integration gets 'Sloppy'

To the editor:

Data integration in life sciences currently faces a conundrum[1–4]. On the one hand, the diversity of data is increasing as explosively as its volume. This makes it imperative that some degree of data formatting standardization is agreed upon by the diverse community generating and using that data. On the other hand, the value of individual data sets can only be appreciated when enough of those distinct pieces of the systemic puzzle are put together. Therefore, it is also imperative that standard formats not be enforced so strictly as to be an obstacle to reporting the very novel data that brings value to the targeted systemic integration. We present here a prototype application, termed Simple Sloppy Semantic Database (S3DB), that provides a bridge between loosely structured raw data annotated using personal ontologies and a globally referenceable semantic representation indexed to controlled vocabularies. Wide adoption of this database formalism has the potential to facilitate and optimize data management in a range of research fields, from molecular epidemiology to basic biology.

For most types of biological data, the agreed-upon communal format has a complexity that is far from trivial and requires specialized converters that were not available when the analytical method was first developed. For example, an agreed-upon Minimum Information about Microarray Experiments (MIAME) standard was defined in 2001 (ref. 5), but the jury is still out for much older and widely used techniques such as gel-based proteomics (for example, see ref. 6). Even when, after much consultation, a community standard emerges, the rigidity of minimal descriptions eventually becomes insufficient for stand-alone reposition[7]. Like many others before us, we have reached the conclusion that complementary efforts in proteomics[8], transcriptomics[9] and genomics[10] can only be integrated in a common representation within a semantic framework[2,11]. We have specifically argued[2] for the need to migrate to RDF (Resource Description Framework) from the more widely used XML (Extensible Markup Language) hierarchies or relational structures, a view also espoused by the World Wide Web consortium Life Sciences interest group (http://www.w3.org/2001/sw/hcls/). However, that formalism is cumbersome for configuring information management systems and trades human intuitiveness for machine process expressiveness. This combination of implementation and interface challenges typically loses the very contribution that is needed to put the systemic puzzle together: that of the 'biology domain' expert.

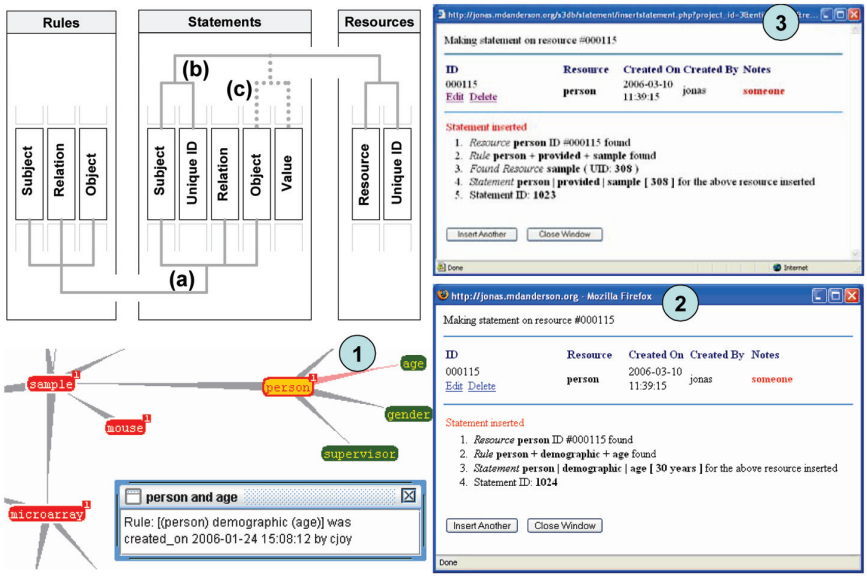

Figure 1 Example of a S3DB application. The indexing scheme is described by the table in the upper left, where the connecting lines identify the three clauses, (a)–(c), verified by the validation engine for a new statement. Three snapshots of the S3DB application for the example discussed in the text are displayed: directed graph depiction of the rules (1), validation log for submission of a literal (nuclear data element such as '30 years') (2) and validation log for the association of two resources (3).

© 2006 Nature Publishing Group http://www.nature.com/naturebiotechnology