BIOLOGICAL
CRYSTALLOGRAPHY

Volume 71 (2015)

Supporting information for article:

The structure of the giant haemoglobin from *Glossoscolex paulistus*

José Fernando Ruggiero Bachega, Fernando Vasconcelos Maluf, Babak Andi, Humberto D'Muniz Pereira, Marcelo Falsarella Carazzollea, Allen M. Orville, Marcel Tabak, José Brandão-Neto, Richard Charles Garratt and Eduardo Horjales Reboredo

S1. Comparison of experimental and theoretical molecular masses

In Supplementary Table 1 the experimental masses obtained by MALDI-TOF-MS analysis [Oliveira et al., 2007; Carvalho et al, 2011] are summarized together with those derived from the sequences reported here. For the linkers L1 and L2 the masses are quite similar. The positive difference observed for the experimental masses could be due to post-translational modifications such as glycosylation, as observed for L1, L2 and L2b. For L3 the mass from the sequence is quite different from that from mass spectrometry. Probably, L3 was not detected in mass spectrometry experiments due to its low concentration, and the mass of 32kDa assigned as due to a mixture of L3 and a dimer of d monomers, corresponds solely to the latter. In the case of chains a and d mass spectral analysis detected up to four isoforms, two of them being predominant. As noticed in Supplementary Table 1, masses of chains a and c were exchanged in the experimental mass assignments. The mass differences (between mass spectrometry and DNA sequencing) for the a isoforms suggest the presence of glycosylation, which has been observed also for the a chains of HbLt [Ownby et al, 1993]. The largest discrepancy is noticed for chain c. The explanation could be due to multiple glycosylations at additional sites, or longer chains at fewer sites. The latter possibility is consistent with disorder as the glycosylation chain length extends further. Disorder is not visible in crystallography. In summary, considering all of the effects described above, the agreement between the directly determined experimental masses and those derived from the deduced amino acid sequences is rather good.

S2. Structure solution

A preliminary report has already described the basic structure solution (Bachega *et al.*, 2011). Briefly, the problem of phase ambiguity and the space group (*I*222 or $I2_12_12_1$)

was solved by molecular replacement using the HbLt protomer (PDB code 2gtl) as template. The search model therefore corresponded to only $1/12^{th}$ of the biological particle from which the haem groups and expected non-haem metals had been excluded. Subsequently, the appearance of density corresponding to these groups was used as an evaluation criterion for the molecular replacement solution. Three protomers were located in the asymmetric unit, a quarter of the biological particle, with two of them coming from one of the hexagonal layers of the biological particle and the third from the other layer of the same particle. The biological unit (with 622 point group symmetry), is therefore generated by the crystallographic symmetry present in space group *I*222. This requires that the centre of the particle sits on a special position, namely the intersection of the three 2-fold axes. The location of the first protomer in the asymmetric unit generated an R value of 57.2% and an LLG of 1,446. The location of the second protomer resulted in an R value of 53.5% and an LLG of 5,509, and subsequently when the third and final protomer was located, the R value dropped to 49.6% and LLG rose significantly to 12,050.

S3. Detailed description of the hetero-trimeric coiled coil formed by the linker chains

The first helix is the larger of the two and corresponds to approximately four heptad repeats making eight sets of hydrophobic contacts (four at the *a* position and four at the *d* position). There is no clear pattern of β-branched and non β-branched residues at the *a* and *d* positions, an observation consistent with the formation of a trimeric coiled-coil over a dimer or tetramer (Harbury et al., 1993). Overall, leucine and isoleucine are the most abundant residues but at the first *d* level two hydrophilic residues, Gln17 and Asn24 are observed in the L1 and L2 chains respectively. This arrangement is stabilized by direct hydrogen bonding which also extends to include Asn15 from the *e* position of the equivalent heptad in the L3 chain (Figure 6). These hydrophilic residues

are conserved in sequences from other species suggesting they represent a common feature. At the C-terminus of the first helix, Arg35 which occupies a *d* position in the L3 chain is charge compensated by the carbonyl groups coming from the remaining two *d* residues, Leu38(L1) and Ala45(L2). This arginine is well conserved in the L3 chains of different species and its role in helix disruption is presumably of structural importance. However, different from *G.paulistus*, the sequence of *A.marina* in this region shows a contiguous pattern of heptads (Fig. 2), consistent with the observed presence of a continuous helix in the crystal structure (Royer *et al.*, 2007). The disruption of the coiled coil may therefore be necessary for determining the relative offset of the protomers from one disc with respect to the other, which is the major structural difference observed between type I particles such as that observed here for *G. paulistus* and type II particles seen in *A. marina*.

Aspartic acid residues are found in all three linkers at the beginning of the second helix but due to variation in the connecting loop these do not occupy equivalent positions in the sequence alignment. In the case of the L2 and L3 chains, these are well conserved and form classical N-caps (Richardson & Richardson, 1988). It is of note, however, that the L3b isoform lacks this aspartic acid (alignment position 61 in Fig 2) which together with other sequence differences in this region suggests possible conformational variation. The second helix is shorter and composed of only two heptad repeats. At the first *a* level, the helices are splayed apart such that a fourth residue (Phe46(L2) occupying a *g* position) contributes to the hydrophobic core. At the C-terminal end, the coiled coil is stabilized by classical Arg (at *g*) – Glu (at *e*) salt bridges.

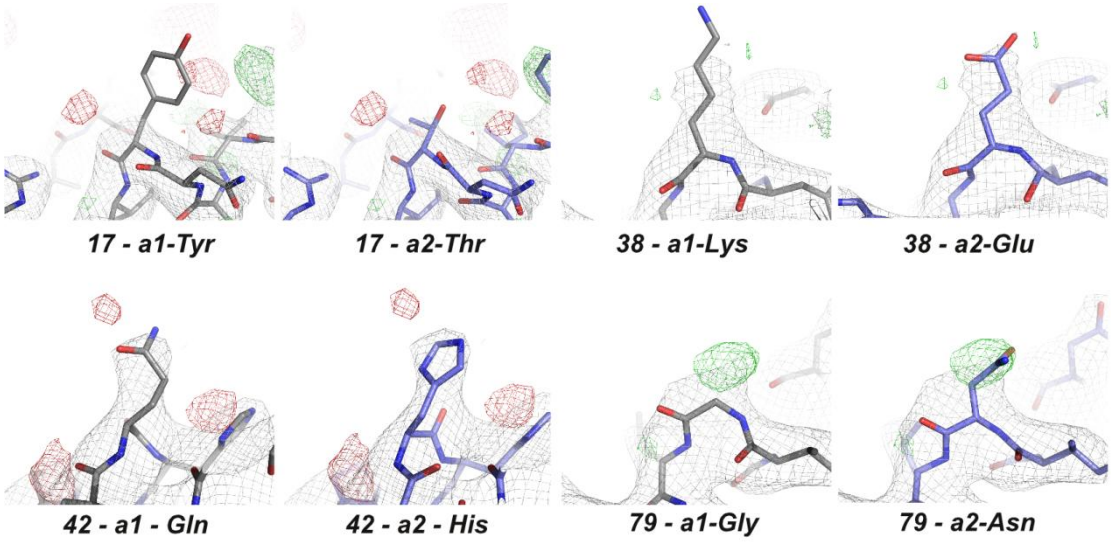

Figure S1 Resolving sequence ambiguity. Omit maps are shown for a series of positions where there is residue variation between the *a1* and *a2* isoforms. In all cases the maps can be seen to be more consistent with the *a2* sequence. Grey density corresponds to the 2*Fo-Fc* map and the green peaks are positive peaks in the difference map.

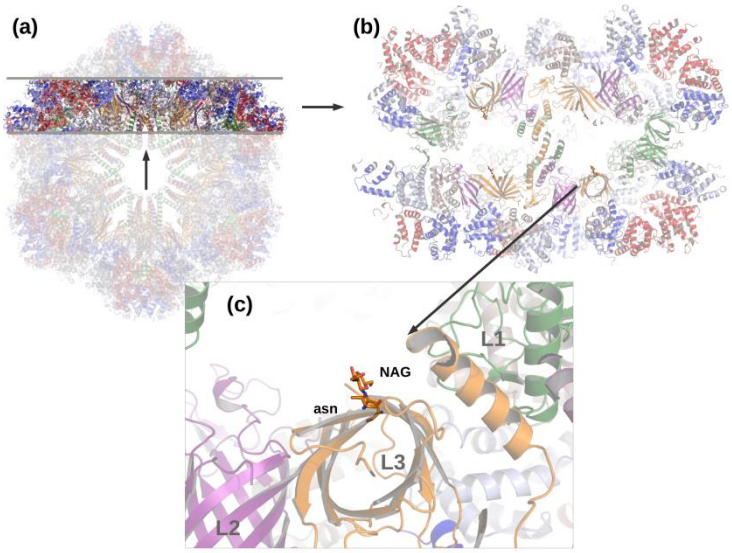

Figure S2

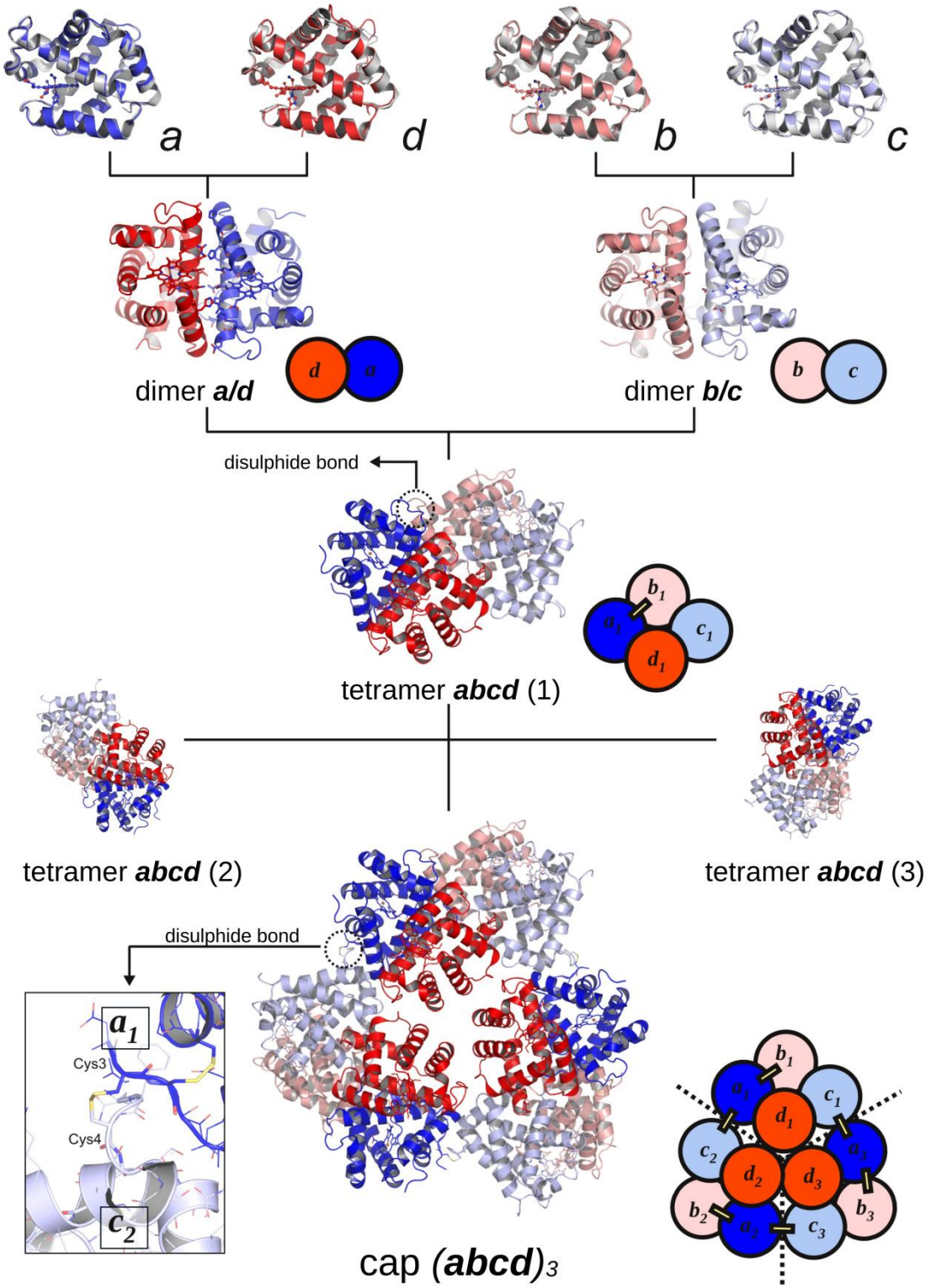

Figure S3 Organization of the dodecameric globin cap**.** *a* and *d* chains and *b* and *c* chains associate via similar interfaces. These dimers pair to form tetramers, three of which associate to form the dodecamer. The simplified representation in the form of spheres follows the same

colour code as the ribbons diagram. Bottom left is shown the electron density for an inter-tetrameric disulphide bridge (boxed).

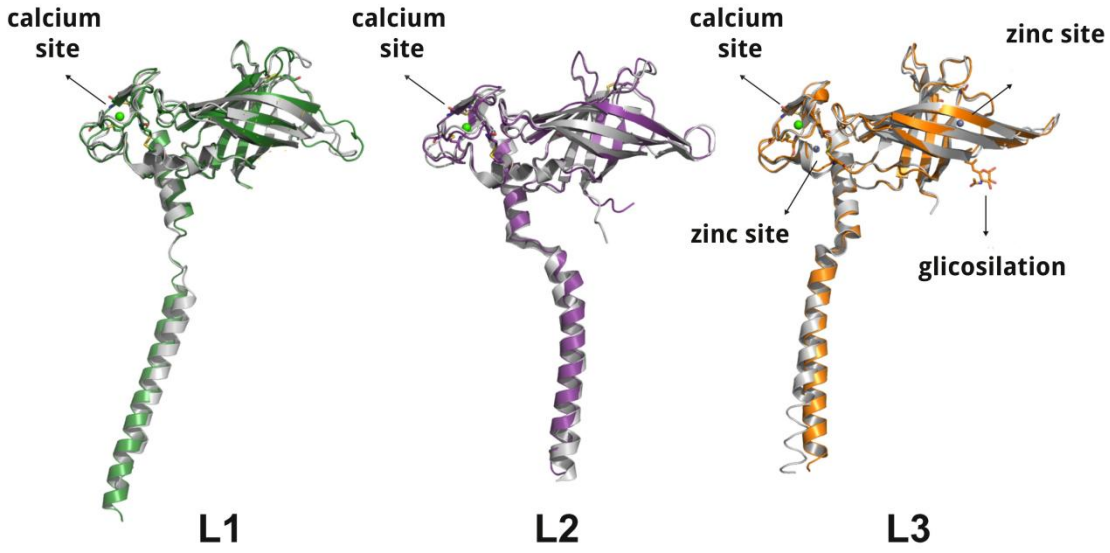

Figure S4 Ribbons representation of the linker chains. The structures from HbGp are shown coloured and superimposed upon their homologs from HbLt (in grey). Overall the folds are very similar with minimal deviations of the long α-helix (domain 1). L3 in HbGp has additional zinc and glycosylation sites not reported previously.

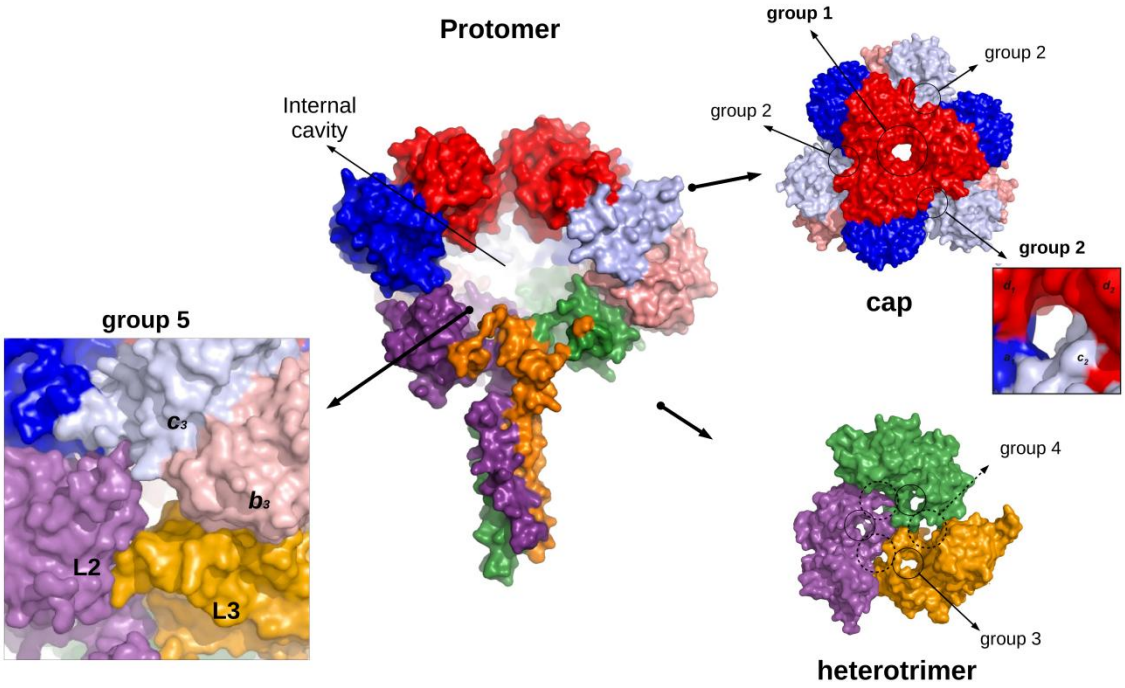

Figure S5 The cavity within the protomer. Defects at inter-domain and inter-subunit interfaces lead to the existence of 5 different types of channel giving access to the central cavity which is lined by the concave face of the cap and the heads of the linkers. The five different types of channel are shown surrounding a protomer (centre) from which part of the structure has been removed to reveal the cavity. Colour coding for the seven different chains is as for other figures.

Table S1 Comparison of predicted physicochemical properties for the 11 HbGp sequences identified in this study

For both molecular mass and pI, where appropriate, the values for *Lumbricus terrestris* haemoglobin are given in parentheses. In terms of predicted pI, the isoforms *d2* and L1b show the most divergent values for the groups to which they belong. The signal peptide sequences were predicted using the SignalP server and show an elevated content of hydrophobic residues as anticipated. The "." indicates the predicted cleavage site.

Sequence	Mass	Experimental mass*	pI	Glicosilation	Signal peptide
globin *a1*	17139,6 (17236,7)	17410[A]	6,39 (7,27)	0	**MKFGLCLVLVAVLGYAFA.DDDCCS**
globin *a2*	17095,2(17236,7)	17330[A]	5,96 (7,27)	0	**FGLCLVLVAVLGYAFA**.DDDCCS
globin *b*	16356,4 (15983,8)	16480	6,85 (5,89)	0	**MKTLVVLLALVAVAMA.**SECDVL
globin *c*	16780,3 (16921,4)	18245[B]	6,32 (6,06)	0	**MLRLLVLLGLAACSMAARA.**HQFCC
globin *d1*	16099,2 (15964,2)	16364[C]	6,19 (5,74)	0	**MKVALVFLLLGVAAFASAD.**DCSILE
globin *d2*	16539,6 (15964,2)	16437[C]	7,22 (5,74)	0	**MKPQIVVFIVLVSYVSGQ.**CSILESLK
Linker L1	25612,4 (25775,5)	25834	5,34 (5,21)	0	**MWYILLLVGLAAAA.**SDTYKDR
Linker L1b	25726,5 (-)	-	7,66 (-)	1	**MRCLGKLLLLLLPVGMA.**DFGS
Linker L2	26729,0 (26064,1)	26785	6,35 (5,51)	1	**MWRLLILSALVGLILA.**DEPQGT
Linker L3	24693,1 (24911,4)	32900[D]	4,83 (4,71)	1	**MKSLRLLLPALAVITAVLA.**DHH
Linker L3b	24518,5 (24113,8)	-	5,27 (5,21)	0	**MRISTRILDAGTLAVLTLVLSFQRVDAA.**EDP

* From MALDI-TOF-MS studies reported in references (International Journal Biological Macromolecules 40(2007) 429-436 and Process Biochemistry 46 (2011) 2144-2151)

A – Assigned in mass spectrometric experiment as subunits *c*, and *two additional isoforms were detected.*

B – Assigned in mass spectrometric experiment as subunits *a*.

C – In mass spectrometric experiments two additional monomeric isoforms were detected.

D – Assigned in mass spectrometric experiments as superposition of linker L3 and dimer of monomers d_x2.